A General MIMO Framework for NOMA Downlink and Uplink Transmission Based on Signal Alignment

Zhiguo Ding, *Senior Member, IEEE*, Robert Schober, *Fellow, IEEE*, and H. Vincent Poor, *Fellow, IEEE*

***Abstract*—The application of multiple-input multiple-output (MIMO) techniques to non-orthogonal multiple access (NOMA) systems is important to enhance the performance gains of NOMA. In this paper, a novel MIMO-NOMA framework for downlink and uplink transmission is proposed by applying the concept of signal alignment. By using stochastic geometry, closed-form analytical results are developed to facilitate the performance evaluation of the proposed framework for randomly deployed users and interferers. The impact of different power allocation strategies, namely fixed power allocation and cognitive radio inspired power allocation, on the performance of MIMO-NOMA is also investigated. Computer simulation results are provided to demonstrate the performance of the proposed framework and the accuracy of the developed analytical results.**

I. Introduction

Non-orthogonal multiple access (NOMA) has been recognized as a spectrally efficient multiple access (MA) technique for the next generation of mobile communication networks [1]–[3]. For example, the use of NOMA has been recently proposed for downlink scenarios in 3rd generation partnership project long-term evolution (3GPP-LTE) systems, and the considered technique was termed multiuser superposition transmission (MUST) [4]. In addition, NOMA has also been identified as one of the key radio access technologies to increase system capacity and reduce latency in fifth generation (5G) mobile networks [5], [6].

The key idea of NOMA is to exploit the power domain for multiple access, which means multiple users can be served concurrently at the same time, frequency, and spreading code. Instead of using water-filling power allocation strategies, NOMA allocates more power to the users with poorer channel conditions, with the aim to facilitate a balanced tradeoff between system throughput and user fairness. Initial system implementations of NOMA in cellular networks have demonstrated the superior spectral efficiency of NOMA [1], [2]. The performance of NOMA in a network with randomly deployed single-antenna nodes has been investigated in [3]. User fairness in the context of NOMA has been addressed in [7], where power allocation was investigated under different channel state information (CSI) assumptions. In [8] and [9], topological interference management has been applied for single-antenna downlink NOMA transmission. Unlike the above works, [10] addressed the application of NOMA for uplink transmission, where the problems of power allocation and subcarrier allocation were jointly considered. The concept of NOMA is not limited to radio frequency communication networks, and has been recently applied to visible light communication systems in [11].

The application of multiple-input multiple-output (MIMO) technologies to NOMA is important since the use of MIMO provides additional degrees of freedom for further performance improvement. The transceiver design for a special case of MIMO-NOMA downlink transmission, in which each user has a single antenna and the base station has multiple antennas, has been investigated in [12] and [13]. In [14], a multiple-antenna base station used the NOMA approach to serve two multiple-antenna users simultaneously, where the problem of throughput maximization was formulated and two algorithms were proposed to solve the optimization problem. In many practical scenarios, it is preferable to serve as many users as possible in order to reduce user latency and improve user fairness. Following this rationale, in [15], users were grouped into small-size clusters, where NOMA was implemented for the users within one cluster and MIMO detection was used to cancel inter-cluster interference. Similar to [16], this method does not need CSI at the base station; however, unlike [16], it avoids the use of random beamforming which can cause uncertainties for the quality of service (QoS) experienced by the users. However, precoding was not considered in [15], which has two disadvantages, as explained in the following. Firstly, in [15], the number of antennas of each user has to be larger than the number of antennas at the base station. Secondly, there is a loss in diversity order since, without precoding, the spatial degrees of freedom at the base station cannot be exploited.

This paper considers a general MIMO-NOMA communication network where a base station is communicating with multiple users using the same time, frequency, and spreading code resources, in the presence of randomly deployed interferers. The contributions of this paper are listed as follows.

Firstly, a general MIMO-NOMA framework which is applicable to both downlink and uplink transmission is proposed, by applying the concept of signal alignment. Recall that signal alignment can be viewed as a special case of interference alignment which can effectively exploit the excess degrees of freedom in MIMO systems for suppressing co-channel interference [17] and [18]. Compared to interference alignment, a key difference of signal alignment is that signals are superimposed, instead of interference. Signal alignment was originally developed for multi-way relaying channels in [19] and [20]. By extending the feature of signal alignment to NOMA, the considered multi-user MIMO-NOMA scenario can be decomposed into multiple separate single-antenna

The material in this paper was presented in part at the IEEE International Conference on Communications, Kuala Lumpur, May 2016.

Z. Ding and H. V. Poor are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA. Z. Ding is also with the School of Computing and Communications, Lancaster University, LA1 4WA, UK. R. Schober is with the Institute for Digital Communications, University of Erlangen-Nurnberg, Germany.

The work of Z. Ding was supported by the UK EPSRC under grant number EP/L025272/1 and by H2020-MSCA-RISE-2015 under grant number 690750. The work of R. Schober was supported by the Alexander von Humboldt Professorship Program. The work of H. V. Poor was supported by National Science Foundation Grant CCF-1420575.

NOMA channels, to which conventional NOMA protocols can be applied straightforwardly.

Secondly, since the choice of the power allocation coefficients is key to achieve a favorable throughput-fairness tradeoff in NOMA systems, two types of power allocation strategies are studied in this paper. The fixed power allocation strategy can realize different QoS requirements in the long term, whereas the cognitive radio inspired power allocation strategy can ensure that the users' QoS requirements are met instantaneously.

Thirdly, a sophisticated approach for the user precoding/detection vector selection is proposed and combined with the signal alignment framework in order to efficiently exploit the excess degrees of freedom of the MIMO system. Compared to the existing MIMO-NOMA work in [15], the framework proposed in this paper offers two benefits. First, unlike the scheme in [15], the proposed scheme is more general and applicable to scenarios where the base station has more antennas than the users. Second, the spatial degrees of freedom of the base station can be efficiently utilized, and a larger diversity gain can be achieved, e.g., for a scenario in which all nodes are equipped with M antennas, a diversity order of M is achievable, whereas a diversity order of 1 is obtained with the scheme in [15]. Numerical results are provided to demonstrate that the MIMO-NOMA scheme proposed in this paper yields a significant performance gain in terms of reception reliability compared to [15].

Finally, exact expressions and asymptotic performance results are developed in order to obtain an insightful understanding of the proposed MIMO-NOMA framework. In particular, the outage probability is used as the performance criterion since it not only bounds the error probability of detection tightly, but also can be used to calculate the outage capacity/rate. The impact of the random locations of the users and the interferers is captured by applying stochastic geometry, and the diversity order is computed to illustrate how efficiently the degrees of freedom of the channels are used by the proposed framework.

II. System Model for the Proposed MIMO-NOMA Framework

Consider an MIMO-NOMA downlink (uplink) communication scenario in which a base station is communicating with multiple users. The base station is equipped with M antennas and each user is equipped with N antennas. In this paper, we consider the scenario $N > \frac{M}{2}$ in order to implement the concept of signal alignment, an assumption more general than the one used in [15]. This assumption is applicable to various communication scenarios, such as small cells in heterogenous networks [21] and 5G cloud radio access (C-RAN) networks [22], in which low-cost base stations are deployed with high density and it is reasonable to assume that the base stations have capabilities similar to those of user handsets, such as smart phones and tablets.

The users are assumed to be uniformly deployed in a disc, denoted by $\mathcal{D}$, i.e., the cell controlled by the base station. The radius of the disc is r, and the base station is located at the center of $\mathcal{D}$. In order to reduce the system load, many existing studies on NOMA have proposed to pair two users for the implementation of NOMA, and have demonstrated that it is ideal to pair users whose channel conditions are very different [1], [23]. Based on this insight, we assume that the disc is divided into two regions as shown in Fig. 1. The first region is a smaller disc, denoted by $\mathcal{D}_1$, with radius r_1 ($r_1 < r$) and the base station located at its origin. The second region is a ring, denoted by $\mathcal{D}_2$, constructed from $\mathcal{D}$ by removing $\mathcal{D}_1$. Assume that M pairs of users are selected, where user m, randomly located in $\mathcal{D}_1$, is paired with user m', randomly located in $\mathcal{D}_2$. Note that there are other types of user pairing strategies, such as random user pairing, which can be used, particularly in the NOMA uplink case, and it is important to point out that the proposed NOMA scheme can be applied if a user pairing strategy different from the one used in this paper. Particularly, as can be seen in the following section, the proposed approach is to manipulate the users' channel matrices and therefore it is applicable even if the distance from the base station to user m is larger than or equal to that from the base station to user m'. The use of more sophisticated user pairing strategies can further improve the performance of the proposed MIMO-NOMA framework of course, but this is beyond the scope of this paper.

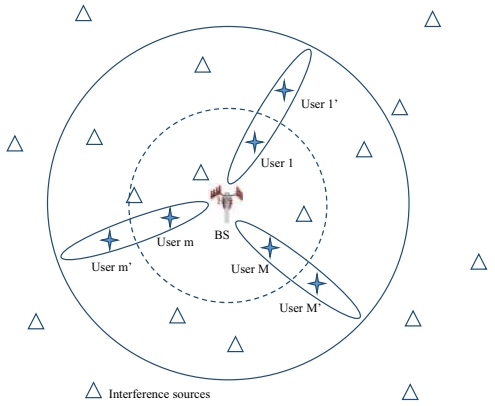


Fig. 1. Illustration of the considered system model.

In addition to the messages sent by the base station, the downlink NOMA users also observe signals sent by interference sources which are distributed in $\mathcal{R}^2$ according to a homogeneous Poisson point process (PPP) Ψ_I of density λ_I [24]. The same assumption is made for the uplink case. In practice, these interferers can be cognitive radio transmitters, WiFi access points in LTE in the unlicensed spectrum (LTE-U), or transmitters from different tiers in heterogenous networks. In order to obtain tractable analytical results, it is assumed that the interference sources are equipped with a single antenna and use identical transmission powers, denoted by ρ_I. These assumptions are valid for cognitive radio networks with single-antenna secondary users. Another example are applications which combine the Internet of Things (IoT) with cellular networks, as the randomly deployed low-cost IoT sensors, which most likely have a single antenna, are potential interference sources for the cellular networks.

Consider the use of a composite channel model with both quasi-static Rayleigh fading and large scale path loss. In

particular, the channel matrix from the base station to user m is $\mathbf{H}_m = \frac{\mathbf{G}_m}{\sqrt{L(d_m)}}$, where $\mathbf{G}_m$ denotes an $N \times M$ matrix whose elements represent Rayleigh fading channel gains, d_m denotes the distance from the base station to the user, and the resulting path loss is modelled as follows:

$$L(d_m) = \begin{cases} d_m^{\alpha}, & \text{if} \quad d_m > r_0 \\ r_0^{\alpha}, & \text{otherwise} \end{cases},$$

where α denotes the path loss exponent and parameter r_0 avoids a singularity when the distance is small. It is assumed that $r_1 \geq r_0$ in order to simplify the analytical results. For notational simplicity, the channel matrix from user m to the base station is denoted by $\mathbf{H}_m^H$. Global CSI is assumed to be available at the users and the base station. The proposed MIMO-NOMA framework for downlink and uplink transmission is described in the following two subsections, respectively.

A. Downlink MIMO-NOMA Transmission

The base station sends the following $M \times 1$ information-bearing vector

$$\mathbf{s} = \begin{bmatrix} \alpha_1 s_1 + \alpha_{1'} s_{1'} \\ \vdots \\ \alpha_M s_M + \alpha_{M'} s_{M'} \end{bmatrix}, \tag{1}$$

where s_m is the signal intended for the m-th user, α_m is the power allocation coefficient, and $\alpha_m^2 + \alpha_{m'}^2 = 1$. The choice of the power allocation coefficients will be discussed later.

Without loss of generality, we focus on user m, whose observation is given by

$$\mathbf{y}_m = \frac{\mathbf{G}_m}{\sqrt{L(d_m)}}\mathbf{P}\mathbf{s} + \mathbf{w}_{I_m} + \mathbf{n}_m, \tag{2}$$

where $\mathbf{P}$ is an $M \times M$ precoding matrix to be defined at the end of this subsection, $\mathbf{w}_{I_m}$ denotes the overall co-channel interference received by user m, and $\mathbf{n}_m$ denotes the noise vector. Following the classical shot noise model in [25], the co-channel interference, $\mathbf{w}_{I_m}$, can be expressed as follows:

$$\mathbf{w}_{I_m} \triangleq \sum_{j \in \Psi_I} \frac{\sqrt{\rho_I}}{\sqrt{L(d_{I_j,m})}} \mathbf{1}_N \tilde{w}_{I_j}, \tag{3}$$

where $\mathbf{1}_m$ denotes an $m \times 1$ all-one vector, ρ_I denotes the interference power, $\tilde{w}_{I_j}$ denotes the normalized signal sent by the interferer, and $d_{I_j,m}$ denotes the distance from user m to the j-th interference source. In this paper, for mathematical tractability, we have omitted small-scale fading in the considered interference model, an assumption commonly used in many existing works to capture the random locations of the users [24]–[27]. This assumption will be discussed more in details in the next section, e.g. Fig. 2. Note that the case with $\rho_I = 0$ corresponds to the scenario without interference.

User m applies a detection vector $\mathbf{v}_m$ to its observation, and therefore the user's observation can be re-written as follows:

$$\begin{aligned} \mathbf{v}_m^H \mathbf{y}_m &= \mathbf{v}_m^H \frac{\mathbf{G}_m}{\sqrt{L(d_m)}}\mathbf{P}\mathbf{s} + \mathbf{v}_m^H(\mathbf{w}_{I_m} + \mathbf{n}_m) \\ &= \mathbf{v}_m^H \frac{\mathbf{G}_m}{\sqrt{L(d_m)}}\mathbf{p}_m(\alpha_m s_m + \alpha_{m'} s_{m'}) \\ &\quad + \underbrace{\sum_{i \neq m} \mathbf{v}_m^H \frac{\mathbf{G}_m}{\sqrt{L(d_m)}}\mathbf{p}_i(\alpha_i s_i + \alpha_{i'} s_{i'}) + \mathbf{v}_m^H(\mathbf{w}_{I_m} + \mathbf{n}_m)}_{interference\ (including\ inter-pair\ interference)\ +\ noise}, \end{aligned} \tag{4}$$

where $\mathbf{p}_m$ denotes the m-th column of $\mathbf{P}$.

In order to remove inter-pair interference, the following constraint has to be met:

$$\begin{bmatrix} \mathbf{v}_m^H \mathbf{G}_m \\ \mathbf{v}_{m'}^H \mathbf{G}_{m'} \end{bmatrix} \mathbf{p}_i = \mathbf{0}_{2\times 1}, \ \forall i \neq m, \tag{5}$$

where $\mathbf{0}_{m\times n}$ denotes the $m \times n$ all zero matrix. Without loss of generality, we focus on $\mathbf{p}_1$ which needs to satisfy the following constraint:

$$\begin{bmatrix} \mathbf{G}_2^H \mathbf{v}_2 & \mathbf{G}_{2'}^H \mathbf{v}_{2'} & \cdots & \mathbf{G}_M^H \mathbf{v}_M & \mathbf{G}_{M'}^H \mathbf{v}_{M'} \end{bmatrix}^H \mathbf{p}_1 = \mathbf{0}. \tag{6}$$

Note that the dimension of the matrix in (6), $\begin{bmatrix} \mathbf{G}_2^H \mathbf{v}_2 & \mathbf{G}_{2'}^H \mathbf{v}_{2'} & \cdots & \mathbf{G}_M^H \mathbf{v}_M & \mathbf{G}_{M'}^H \mathbf{v}_{M'} \end{bmatrix}^H$, is $2(M-1) \times M$. Therefore, in general, a non-zero vector $\mathbf{p}_i$ satisfying (6) does not exist. In order to ensure the existence of $\mathbf{p}_i$, one straightforward approach is to serve less user pairs, i.e., reducing the number of user pairs to $\left(\frac{M}{2} + 1\right)$. However, this approach will reduce the overall system throughput.

To overcome this problem, in this paper, the concept of signal alignment is applied, which means the detection vectors are designed to satisfy the following constraint [28], [29]

$$\mathbf{v}_m^H \mathbf{G}_m = \mathbf{v}_{m'}^H \mathbf{G}_{m'}, \tag{7}$$

or equivalently $\begin{bmatrix} \mathbf{G}_m^H & -\mathbf{G}_{m'}^H \end{bmatrix} \begin{bmatrix} \mathbf{v}_m \\ \mathbf{v}_{m'} \end{bmatrix} = \mathbf{0}_{M\times 1}$. Define $\mathbf{U}_m$ as the $2N \times (2N - M)$ matrix containing the $(2N - M)$ right singular vectors of $\begin{bmatrix} \mathbf{G}_m^H & -\mathbf{G}_{m'}^H \end{bmatrix}$ corresponding to its zero singular values. Therefore, the detection vectors at the users are designed as follows:

$$\begin{bmatrix} \mathbf{v}_m \\ \mathbf{v}_{m'} \end{bmatrix} = \mathbf{U}_m \mathbf{x}_m, \tag{8}$$

where $\mathbf{x}_m$ is a $(2N - M) \times 1$ vector to be defined later. We normalize $\mathbf{x}_m$ to 2, i.e., $|\mathbf{x}|^2 = 2$, due to the following two reasons. First, the uplink transmission power has to be constrained as shown in the following subsection. Second, this facilitates the performance analysis carried out in the next section. It is straightforward to show that the choice of the detection vectors in (8) satisfies $\begin{bmatrix} \mathbf{G}_m^H & -\mathbf{G}_{m'}^H \end{bmatrix} \mathbf{U}_m \mathbf{x}_m = \mathbf{0}_{M\times 1}$.

The effect of the signal alignment based design in (7) is the projection of the channels of the two users in the same pair into the same direction. Define $\mathbf{g}_m \triangleq \mathbf{G}_m^H \mathbf{v}_m$ as the effective channel vector shared by the two users. As a result, the number of rows in the matrix in (6) can be reduced significantly. In particular, the constraint for $\mathbf{p}_i$ in (6) can be rewritten as follows:

$$\begin{bmatrix} \mathbf{g}_1 & \cdots & \mathbf{g}_{i-1} & \mathbf{g}_{i+1} & \cdots & \mathbf{g}_M \end{bmatrix}^H \mathbf{p}_i = \mathbf{0}_{(M-1)\times 1}. \tag{9}$$

Note that $\begin{bmatrix}\mathbf{g}_1 & \cdots & \mathbf{g}_{i-1} & \mathbf{g}_{i+1} & \cdots & \mathbf{g}_M\end{bmatrix}^H$ is an $(M-1)\times M$ matrix, which means that a $\mathbf{p}_i$ satisfying (9) exists.

Define $\mathbf{G} \triangleq \begin{bmatrix}\mathbf{g}_1 & \cdots & \mathbf{g}_M\end{bmatrix}^H$. A zero forcing based precoding matrix at the base station can be designed as follows:

$$\mathbf{P} = \mathbf{G}^{-H}\mathbf{D}, \tag{10}$$

where $\mathbf{D}$ is a diagonal matrix to ensure power normalization at the base station, i.e., $\mathbf{D}^2 = \text{diag}\{\frac{1}{(\mathbf{G}^{-1}\mathbf{G}^{-H})_{1,1}}, \cdots, \frac{1}{(\mathbf{G}^{-1}\mathbf{G}^{-H})_{M,M}}\}$, where $(\mathbf{A})_{m,m}$ denotes the m-th element on the main diagonal of $\mathbf{A}$. As a result, the transmission power at the base station can be constrained as follows:

$$\text{tr}\left\{\mathbf{P}\mathbf{P}^H\right\}\rho = \text{tr}\left\{\mathbf{G}^{-1}\mathbf{G}^{-H}\mathbf{D}^2\right\}\rho = M\rho, \tag{11}$$

where ρ denotes the transmit signal-to-noise ratio (SNR).

With the design in (7) and (10), the signal model for user m can now be written as follows:

$$\mathbf{v}_m^H\mathbf{y}_m = \frac{(\alpha_m s_m + \alpha_{m'}s_{m'})}{\sqrt{(L(d_m))(\mathbf{G}^{-1}\mathbf{G}^{-H})_{m,m}}} + \mathbf{v}_m^H(\mathbf{w}_{I_m} + \mathbf{n}_m). \tag{12}$$

For notational simplicity, we define $y_m = \mathbf{v}_m^H\mathbf{y}_m$, $h_m = \frac{1}{\sqrt{L(d_m)(\mathbf{G}^{-1}\mathbf{G}^{-H})_{m,m}}}$, $w_{I_m} = \mathbf{v}_m^H\mathbf{w}_{I_m}$, and $n_m = \mathbf{v}_m^H\mathbf{n}_m$. Therefore, the use of the signal alignment based precoding and detection matrices decomposes the multi-user MIMO-NOMA channels into M pairs of single-antenna NOMA channels. In particular, within each pair, the two users receive the following scalar observations

$$y_m = h_m(\alpha_m s_m + \alpha_{m'}s_{m'}) + w_{I_m} + n_m, \tag{13}$$

and

$$y_{m'} = h_{m'}(\alpha_m s_m + \alpha_{m'}s_{m'}) + w_{I_{m'}} + n_{m'}, \tag{14}$$

where $y_{m'}$ and $n_{m'}$ are defined similar to y_m and n_m, respectively. Note that $h_{m'} = \frac{1}{\sqrt{L(d_{m'})(\mathbf{G}^{-1}\mathbf{G}^{-H})_{m,m}}}$. Hence, the two effective channel gains, h_m and $h_{m'}$, share the same small scale fading gain, $\frac{1}{(\mathbf{G}^{-1}\mathbf{G}^{-H})_{m,m}}$, but correspond to different distances.

Recall that two users belonging to the same pair are selected from $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively, which means that $d_m < d_{m'}$. Therefore, the two users from the same pair are ordered without any ambiguity, which simplifies the design of the power allocation coefficients, i.e., $\alpha_m \leq \alpha_{m'}$, following the NOMA principle. User m' decodes its message with the following signal-to-interference-plus-noise ratio (SINR)

$$SINR_{m'} = \frac{\rho|h_{m'}|^2\alpha_{m'}^2}{\rho|h_{m'}|^2\alpha_m^2 + |\mathbf{v}_{m'}|^2 + |\mathbf{v}_{m'}^H\mathbf{1}_N|^2 I_{m'}}, \tag{15}$$

where the interference term is given by

$$I_{m'} = \sum_{j\in\Psi_I}\frac{\rho_I}{L\left(d_{I_j,m'}\right)}. \tag{16}$$

User m carries out successive interference cancellation (SIC) by first removing the message to user m' with SINR, $SINR_{m,m'} = \frac{\rho|h_m|^2\alpha_{m'}^2}{\rho|h_m|^2\alpha_m^2 + |\mathbf{v}_m|^2 + |\mathbf{v}_m^H\mathbf{1}_N|^2 I_m}$, and then decoding its own message with SINR

$$SINR_m = \frac{\rho|h_m|^2\alpha_m^2}{|\mathbf{v}_m|^2 + |\mathbf{v}_m^H\mathbf{1}_N|^2 I_m}, \tag{17}$$

which becomes the SNR if $\rho_I = 0$.

Example 1: Consider a simple example for the proposed signal alignment approach. Assume that there are three user pairs, i.e., $M = 3$, and each user has two antennas, i.e., $N = 2$. Suppose that the instantaneous realizations of the channel matrices for the users in pair 2 and 3 are given by

$$\mathbf{G}_2 = \begin{bmatrix}1 & 0 & 0\\ 1 & 0 & 1\end{bmatrix}, \mathbf{G}_{2'} = \begin{bmatrix}0 & 1 & 0\\ 0 & -1 & 1\end{bmatrix}, \tag{18}$$
$$\mathbf{G}_3 = \begin{bmatrix}1 & 1 & 0\\ 1 & 0 & 0\end{bmatrix}, \text{ and } \mathbf{G}_{3'} = \begin{bmatrix}0 & 1 & 1\\ 0 & 0 & 1\end{bmatrix}.$$

According to (6), the 3×1 precoding vector, $\mathbf{p}_1$, needs to be orthogonal to the 4×3 matrix $\begin{bmatrix}\mathbf{G}_2^H\mathbf{v}_2 & \mathbf{G}_{2'}^H\mathbf{v}_{2'} & \mathbf{G}_3^H\mathbf{v}_3 & \mathbf{G}_{3'}^H\mathbf{v}_{3'}\end{bmatrix}^H$, whose null space does not exist if it is full column rank. Following the exposition in this section, the detection vectors can be designed as $\mathbf{v}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}-1 & 1\end{bmatrix}$, $\mathbf{v}_{2'} = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\end{bmatrix}$, $\mathbf{v}_3 = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & -1\end{bmatrix}$, and $\mathbf{v}_{3'} = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & -1\end{bmatrix}$. With these choices of $\mathbf{v}_i$, $\mathbf{p}_1$ needs to be orthogonal to the following matrix:

$$\begin{bmatrix}\mathbf{G}_2^H\mathbf{v}_2 & \mathbf{G}_{2'}^H\mathbf{v}_{2'} & \mathbf{G}_3^H\mathbf{v}_3 & \mathbf{G}_{3'}^H\mathbf{v}_{3'}\end{bmatrix}^H = \frac{1}{\sqrt{2}}\begin{bmatrix}0 & 0 & 0 & 0\\ 0 & 0 & 1 & 1\\ 1 & 1 & 0 & 0\end{bmatrix}^H. \tag{19}$$

Note that, because of the applied signal alignment, the above matrix has rank 2, although its dimension is 4×3. Therefore, the null space of the matrix exists. In particular, an ideal choice for $\mathbf{p}_1$ is $\mathbf{p}_1 = \begin{bmatrix}1 & 0 & 0\end{bmatrix}^H$, which is orthogonal to the matrix in (19).

B. Uplink MIMO-NOMA Transmission

For the NOMA uplink case, user m will send out an information bearing message s_m, and the signal transmitted by this user is denoted by $\alpha_m\mathbf{v}_m s_m$. Because of the reciprocity between uplink and downlink channels, $\mathbf{v}_m$ which was used as a downlink detection vector can be used as a precoding vector for the uplink scenario. Similarly $\mathbf{P}$ will be used as the detection matrix for the uplink case. In this paper, we assume that the total transmission power from one user pair is normalized as follows:

$$\alpha_m^2|\mathbf{v}_m|^2 + \alpha_{m'}^2|\mathbf{v}_{m'}|^2 \leq 2\rho. \tag{20}$$

The base station observes the following signal:

$$\mathbf{y}_{BS} = \sum_{m=1}^{M}\left(\frac{\mathbf{G}_m^H\alpha_m\mathbf{v}_m s_m}{\sqrt{L(d_m)}} + \frac{\mathbf{G}_{m'}^H\alpha_{m'}\mathbf{v}_{m'}s_{m'}}{\sqrt{L(d_{m'})}}\right) + \mathbf{w}_I + \mathbf{n}_{BS}, \tag{21}$$

where $\mathbf{w}_I$ is the interference term defined as follows

$$\mathbf{w}_I \triangleq \sum_{j\in\Psi_I}\frac{\sqrt{\rho_I}}{\sqrt{L\left(d_{I_j,BS}\right)}}\mathbf{1}_M\tilde{w}_{I_j}, \tag{22}$$

$d_{I_j,BS}$ denotes the distance between the base station and the j-th interferer, and the noise term is defined similarly as in the previous section. The base station applies a detection matrix

$\mathbf{P}$ to its observations and the system model at the base station can be written as follows:

$$\mathbf{P}^H\mathbf{y}_{BS} = \mathbf{P}^H \sum_{m=1}^{M} \left(\frac{\mathbf{G}_m^H \alpha_m \mathbf{v}_m s_m}{\sqrt{L(d_m)}} + \frac{\mathbf{G}_{m'}^H \alpha_{m'} \mathbf{v}_{m'} s_{m'}}{\sqrt{L(d_{m'})}} \right) + \mathbf{P}^H(\mathbf{w}_I + \mathbf{n}_{BS}).$$

As a result, the symbols from the m-th user pair can be detected based on

$$\mathbf{p}_m^H\mathbf{y}_{BS} = \mathbf{p}_m^H \left(\frac{\mathbf{G}_m^H \alpha_m \mathbf{v}_m s_m}{\sqrt{L(d_m)}} + \frac{\mathbf{G}_{m'}^H \alpha_{m'} \mathbf{v}_{m'} s_{m'}}{\sqrt{L(d_{m'})}} \right) + \underbrace{\mathbf{p}_m^H \sum_{i\neq m} \left(\frac{\mathbf{G}_i^H \alpha_i \mathbf{v}_i s_i}{\sqrt{L(d_i)}} + \frac{\mathbf{G}_{i'}^H \alpha_{i'} \mathbf{v}_{i'} s_{i'}}{\sqrt{L(d_{i'})}} \right) + \mathbf{p}_m^H(\mathbf{w}_I + \mathbf{n}_{BS})}_{\text{interference (including inter-pair interference) + noise}}.$$

In order to avoid inter-pair interference, the following constraint needs to be met

$$\mathbf{p}_m^H \sum_{i\neq m} \left(\frac{\mathbf{G}_i^H \alpha_i \mathbf{v}_i s_i}{\sqrt{L(d_i)}} + \frac{\mathbf{G}_{i'}^H \alpha_{i'} \mathbf{v}_{i'} s_{i'}}{\sqrt{L(d_{i'})}} \right) = 0, \ \forall m \neq i. \quad (23)$$

Applying again the concept of signal alignment, the constraint that $\mathbf{G}_m^H\mathbf{v}_m = \mathbf{G}_{m'}^H\mathbf{v}_{m'}$ is imposed on the precoding vectors $\mathbf{v}_m$. Therefore, the same design of $\mathbf{v}_m$ as shown in (8) can be used. The total transmission power within one pair is given by

$$\begin{aligned} &\rho\alpha_m^2|\mathbf{v}_m|^2 + \rho\alpha_{m'}^2|\mathbf{v}_{m'}|^2 \\ &\leq \rho\max(\alpha_m^2, \alpha_{m'}^2)(|\mathbf{v}_m|^2 + |\mathbf{v}_{m'}|^2) \leq 2\rho. \end{aligned} \quad (24)$$

Therefore, the use of the precoding vector in (8) ensures that the total transmission power of one user pair is constrained.

Applying the detection matrix defined in (10), the system model for the base station to decode the messages from the m-th pair can be written as follows:

$$y_{BS,m} = h_m\alpha_m s_m + h_{m'}\alpha_{m'}s_{m'} + w_{BS,m} + n_{BS,m}, \quad (25)$$

where $y_{BS,m} = \mathbf{p}_m^H\mathbf{y}_{BS}$, $w_{BS,m} = \mathbf{p}_m^H\mathbf{w}_I$, and $n_{BS,m} = \mathbf{p}_m^H\mathbf{n}_{BS}$. Therefore, using the proposed precoding and detection matrices, we can decompose the multi-user MIMO-NOMA uplink channel into M orthogonal single-antenna NOMA channels. Note that the variance of the noise is normalized as illustrated in the following:

$$\begin{aligned} &\mathcal{E}\{\mathbf{p}_m^H\mathbf{n}_{BS}\mathbf{n}_{BS}^H\mathbf{p}_m\} = \mathbf{p}_m^H\mathbf{p}_m = (\mathbf{P}^H\mathbf{P})_{m,m} \\ &= (\mathbf{D}^H\mathbf{G}^{-1}\mathbf{G}^{-H}\mathbf{D})_{m,m} = \frac{(\mathbf{G}^{-1}\mathbf{G}^{-H})_{m,m}}{(\mathbf{G}^{-1}\mathbf{G}^{-H})_{m,m}} = 1. \end{aligned} \quad (26)$$

If the message of user m is decoded first, the SINR at the base station to detect this message can be expressed as follows:

$$SINR_{m,BS} = \frac{\rho|h_m|^2\alpha_m^2}{\rho|h_{m'}|^2\alpha_{m'}^2 + I_{BS,m} + 1}, \quad (27)$$

and the SINR for detecting the message of user m' is $SINR_{m',BS} = \frac{\rho|h_{m'}|^2\alpha_{m'}^2}{I_{BS,m}+1}$, where the interference power is related to detection vector $\mathbf{p}_m$ as follows:

$$I_{BS,m} = \sum_{j\in\Psi_I} \frac{\rho_I|\mathbf{p}_m^H\mathbf{1}_M|^2}{L\left(d_{I_j,BS}\right)}. \quad (28)$$

III. Performance Analysis for Downlink MIMO-NOMA Transmission

Two types of power allocation policies are considered in this section. One is fixed power allocation and the other one is inspired by the cognitive ratio concept, as illustrated in the following two subsections, respectively. Recall that the precoding vectors $\mathbf{v}_m$ and $\mathbf{v}_{m'}$ are determined by $\mathbf{x}_m$ as shown in (8). In this section, a random choice of $\mathbf{x}_m$ is considered first. How to find a more sophisticated choice for $\mathbf{x}_m$ is investigated in Section III-C.

A. Fixed Power Allocation

In this case, the power allocation coefficients α_m and $\alpha_{m'}$ are constant and not related to the instantaneous realizations of the fading channels. We will first focus on the outage performance of user m'. The outage probability of user m' to decode its information is given by

$$\mathrm{P}_{m'}^o = \mathrm{P}\left(\log\left(1 + \frac{\rho|h_{m'}|^2\alpha_{m'}^2}{\rho|h_{m'}|^2\alpha_m^2 + |\mathbf{v}_{m'}|^2 + |\mathbf{v}_{m'}^H\mathbf{1}_N|^2 I_{m'}}\right) < R_{m'}\right), \quad (29)$$

where $\mathrm{P}(x<a)$ denotes the probability for the event $x<a$. The correlation between $\mathbf{v}_{m'}$ and $h_{m'}$ makes the evaluation of the above outage probability very challenging. Hence, we focus on the following modified expression for the outage probability

$$\tilde{\mathrm{P}}_{m'} = \mathrm{P}\left(\log\left(1 + \frac{\rho|h_{m'}|^2\alpha_{m'}^2}{\rho|h_{m'}|^2\alpha_m^2 + 2 + 2\delta I_{m'}}\right) < R_{m'}\right). \quad (30)$$

Since $|\mathbf{v}_{m'}|^2 + |\mathbf{v}_m|^2 = 2$, we have $|\mathbf{v}_{m'}|^2 \leq 2$ and $|\mathbf{v}_m|^2 \leq 2$. In addition, because $(\frac{1}{N}\sum_{n=1}^N x_n)^2 \leq \frac{1}{N}\sum_{n=1}^N x_n^2$, $|\mathbf{v}_{m'}^H\mathbf{1}_N|^2 \leq N|\mathbf{v}_{m'}|^2$. Therefore, we have

$$\mathrm{P}_{m'}^o \leq \tilde{\mathrm{P}}_{m'}, \quad (31)$$

for $\delta \geq N$, which means that $\tilde{\mathrm{P}}_{m'}$ provides an upper bound on $\mathrm{P}_{m'}^o$ if $\delta \geq N$. Note that when $\delta = 1$, the difference between $\tilde{\mathrm{P}}_{m'}$ and $\mathrm{P}_{m'}^o$ is very small as can be observed from Fig. 2, i.e., a choice of $\delta = 1$ is sufficient to ensure that $\tilde{\mathrm{P}}_{m'}$ provides a very tight approximation to $\mathrm{P}_{m'}^o$. In addition, the use of $\tilde{\mathrm{P}}_{m'}$ will be sufficient to identify the achievable diversity order of the proposed MIMO-NOMA scheme. These conclusions are valid even if small scale fading is included in the interference model, as can been observed from Fig. 2. It is also worth pointing out that the omission of small scale fading leads to an overestimation of the effect of co-channel interference, as can be seen in Fig. 2. This is because the absolute square of a complex Gaussian variable with zero mean and variance one is likely smaller than one. The figure also reveals that when the interference power is small, the difference between the cases with and without small scale fading becomes negligible.

Given a random choice of $\mathbf{x}_m$, the following lemma provides an exact expression for $\tilde{\mathrm{P}}_{m'}$ as well as its high SNR approximation.

Lemma 1. *If $\alpha_{m'}^2 \leq \alpha_m^2\epsilon_{m'}$, the probability $\tilde{\mathrm{P}}_{m'} = 1$, where $\epsilon_{m'} = 2^{R_{m'}} - 1$. Otherwise the probability $\tilde{\mathrm{P}}_{m'}$ can be expressed as follows:*

$$\tilde{\mathrm{P}}_{m'} = 1 - \frac{2}{r^2 - r_1^2}\int_{r_1}^{r} e^{-2\phi_{m'}x^\alpha}\varphi_I(x)xdx, \quad (32)$$

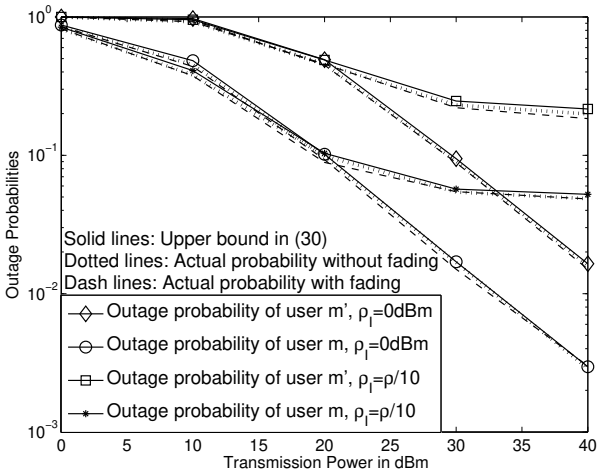


Fig. 2. Comparison between $\tilde{\mathrm{P}}_{i'}$ and $\mathrm{P}^o_{i'}$, $i \in \{m, m'\}$. $R_m = R_{m'} = 1.5$ bit per channel use (BPCU). $\lambda_I = 10^{-4}$. $r = 20$m and $r_1 = 10$m. $r_0 = 1$m and $\alpha_{m'} = \frac{3}{4}$. $M = N = 2$. The path loss exponent is $\alpha = 3$, and the noise power is -30dBm.

where $\phi_{m'} = \frac{\epsilon_{m'}}{\rho\alpha^2_{m'} - \rho\alpha^2_m\epsilon_{m'}}$, $\varphi_I(x) = e^{-\pi\lambda_I(\beta_{m'}(x))^{\frac{2}{\alpha}}\gamma\left(\frac{1}{\alpha}, \frac{\beta_{m'}(x)}{r_0^\alpha}\right)}$, $\beta_{m'}(x) = 2\phi_{m'}\delta\rho_I L\left(x^\alpha\right)$, *and* $\gamma(\cdot)$ *denotes the incomplete Gamma function.*

If ρ_I *is fixed and transmit SNR* ρ *approaches infinity, the outage probability can be approximated as follows:*

$$\tilde{\mathrm{P}}_{m'} \approx \frac{2\phi_{m'}(2+\tilde{\theta}_{m'})}{r^2 - r_1^2}\frac{\left(r^{\alpha+2} - r_1^{\alpha+2}\right)}{\alpha+2}, \tag{33}$$

where $\tilde{\theta}_{m'} = 2\pi\lambda_I\delta\rho_I\frac{\alpha}{r_0}$. *For the special case of* $\rho_I = 0$, $\tilde{\mathrm{P}}_{m'}$ *simplifies to*

$$\tilde{\mathrm{P}}_{m'} = 1 - \frac{1}{r^2 - r_1^2}\left(e^{-2\phi_{m'}r^\alpha}r^2 - e^{-2\phi_{m'}r_1^\alpha}r_1^2\right) - \frac{(2\phi_{m'})^{-\frac{2}{\alpha}}}{r^2 - r_1^2}\left(\gamma\left(\frac{2}{\alpha}+1, 2\phi_{m'}r^\alpha\right) - \gamma\left(\frac{2}{\alpha}+1, 2\phi_{m'}r_1^\alpha\right)\right). \tag{34}$$

Proof. Please refer to Appendix A. □

By using the high SNR approximation obtained in Lemma 1 and also the fact that both $\phi_{m'}$ and $\theta_{m'}$ are proportional to $\frac{1}{\rho}$, the achievable diversity order is obtained in the following corollary.

Corollary 1. *If* $\alpha^2_{m'} > \alpha^2_m\epsilon_{m'}$, *the diversity order achieved by the proposed MIMO-NOMA framework for user* m' *is one.*

On the other hand, user m first decodes the message for user m' before decoding its own message via SIC. Therefore, the outage probability at user m is given by

$$\mathrm{P}^o_m = \mathrm{P}\left(\log\left(1 + \frac{\rho|h_m|^2\alpha^2_{m'}}{\rho|h_m|^2\alpha^2_m + |\mathbf{v}_m|^2 + |\mathbf{v}_m^H 1_N|^2 I_m}\right) < R_{m'}\right) + \mathrm{P}\left(\log\left(1 + \frac{\rho|h_m|^2\alpha^2_m}{|\mathbf{v}_m|^2 + |\mathbf{v}_m^H 1_N|^2 I_m}\right) < R_m, \log\left(1 + \frac{\rho|h_m|^2\alpha^2_{m'}}{\rho|h_m|^2\alpha^2_m + |\mathbf{v}_m|^2 + |\mathbf{v}_m^H 1_N|^2 I_m}\right) > R_{m'}\right). \tag{35}$$

Again, we focus on a modified expression for the outage probability as follows:

$$\tilde{\mathrm{P}}_m = \mathrm{P}\left(\log\left(1 + \frac{\rho|h_m|^2\alpha^2_{m'}}{\rho|h_m|^2\alpha^2_m + 2 + 2\delta I_m}\right) < R_{m'}\right) + \mathrm{P}\left(\log\left(1 + \frac{\rho|h_m|^2\alpha^2_m}{2 + 2\delta I_m}\right) < R_m, \log\left(1 + \frac{\rho|h_m|^2\alpha^2_{m'}}{\rho|h_m|^2\alpha^2_m + 2 + 2\delta I_m}\right) > R_{m'}\right), \tag{36}$$

which is an upper bound for $\delta \geq N$ as explained in the proof for Lemma 2. Fig. 2 demonstrates that $\tilde{\mathrm{P}}_m$ with a choice of $\delta = 1$ yields a tight approximation on P_m. The following lemma provides an exact expression for this probability as well as its high SNR approximation.

Lemma 2. *If* $\alpha^2_{m'} \leq \alpha^2_m\epsilon_{m'}$, *the probability* $\tilde{\mathrm{P}}_m = 1$, *otherwise the probability* $\tilde{\mathrm{P}}_{m'}$ *can be expressed as follows:*

$$\tilde{\mathrm{P}}_m = 1 - \frac{2}{r_1^2}\int_0^{r_0} e^{-2\tilde{\phi}_m r_0^\alpha}\varphi_I(r_0)xdx - \frac{2}{r_1^2}\int_{r_0}^{r_1} e^{-2\tilde{\phi}_m x^\alpha}\varphi_I(x)xdx, \tag{37}$$

where $\tilde{\phi}_m = \max\{\phi_m, \phi_{m'}\}$ *and* $\phi_m = \frac{\epsilon_m}{\rho\alpha^2_m}$. *If* ρ_I *is fixed and the transmit SNR* ρ *approaches infinity, the outage probability can be approximated as follows:*

$$\tilde{\mathrm{P}}_m \approx \frac{\tilde{\phi}_m(2+\tilde{\theta}_{m'})}{r_1^2(\alpha+2)}\left(\alpha r_0^{\alpha+2} + 2r_1^{\alpha+2}\right), \tag{38}$$

where $\tilde{\theta}_{m'}$ *was defined in Lemma 1.*

Proof. Please refer to Appendix B. □

Following steps similar to those in the proof for Corollary 1, one can conclude that the diversity order of user m is the same as that of user m', when $\alpha^2_{m'} > \alpha^2_m\epsilon_{m'}$. This is due to the fact that h_m and $h_{m'}$ share the same small scale fading gain and diff only in the path loss attenuation.

B. Cognitive Radio Power Allocation

In this section, a cognitive radio inspired power allocation strategy is studied. In particular, assume that user m' is viewed as a primary user in a cognitive ratio network. With orthogonal multiple access, the bandwidth resource occupied by user m' cannot be reused by other users, despite its poor channel conditions. In contrast, with NOMA, one additional user, i.e., user m, can be served simultaneously, under the condition that the QoS requirements of user m' can still be met.

In particular, assume that user m' needs to achieve a target data rate of $R_{m'}$, which means that the power allocation coefficients of NOMA need to satisfy the following constraint

$$\frac{\rho|h_{m'}|^2\alpha^2_{m'}}{\rho|h_{m'}|^2\alpha^2_m + |\mathbf{v}_{m'}|^2 + |\mathbf{v}^H_{m'}\mathbf{1}_N|^2 I_{m'}} > \epsilon_{m'}, \tag{39}$$

which leads to the following choice for α_m

$$\alpha^2_m = \max\left(0, \frac{\rho|h_{m'}|^2 - \epsilon_{m'}(|\mathbf{v}_{m'}|^2 + |\mathbf{v}^H_{m'}\mathbf{1}_N|^2 I_{m'})}{(1+\epsilon_{m'})\rho|h_{m'}|^2}\right). \tag{40}$$

It is straightforward to show that $\frac{\rho|h_{m'}|^2-\epsilon_{m'}(|\mathbf{v}_{m'}|^2+|\mathbf{v}_{m'}^H\mathbf{1}_N|^2 I_{m'})}{(1+\epsilon_{m'})\rho|h_{m'}|^2}$ is always less than one.

An outage at user m' means here that all power is allocated to user m', but outage still occurs. As a result, the outage probability of user m' is exactly the same as that in conventional orthogonal multiple access systems. Therefore, in this section, we only focus on the outage probability of user m which can be expressed as follows:

$$\begin{aligned}\mathrm{P}_m^o =&\mathrm{P}\left(|h_m|^2 < \max\left\{\phi_{m'}(|\mathbf{v}_m|^2+|\mathbf{v}_m^H\mathbf{1}_N|^2 I_m),\right.\right. \\ &\left.\left.\phi_m(|\mathbf{v}_m|^2+|\mathbf{v}_m^H\mathbf{1}_N|^2 I_m)\right\}\right),\end{aligned} \tag{41}$$

if $\alpha_{m'}^2 > \alpha_m^2\epsilon_{m'}$, otherwise outage always occurs. It can be verified that $\alpha_{m'}^2 \leq \alpha_m^2\epsilon_{m'}$ is equivalent to $\alpha_m = 0$, in the context of cognitive radio power allocation.

Analyzing the outage probability in (41) is very difficult due to the following two reasons. First, h_m and $\mathbf{v}_m$ are correlated, and second, the users experience different but correlated co-channel interference, i.e., $I_m \neq I_{m'}$. Therefore, in this subsection, we only focus on the case without co-channel interference, i.e., $\rho_I = 0$. In particular, we focus on the following outage probability

$$\tilde{\mathrm{P}}_m =\mathrm{P}\left(|h_m|^2 < 2\max\left\{\bar{\phi}_{m'},\bar{\phi}_m\right\}\right), \tag{42}$$

where $\bar{\phi}_m = \frac{\epsilon_m}{\rho\bar{\alpha}_m^2}$, $\bar{\phi}_{m'} = \frac{\epsilon_{m'}}{\rho\bar{\alpha}_{m'}^2-\rho\bar{\alpha}_m^2\epsilon_{m'}}$, and $\bar{\alpha}_m^2 = \max\left(0, \frac{\rho|h_{m'}|^2-2\epsilon_{m'}}{(1+\epsilon_{m'})\rho|h_{m'}|^2}\right)$.

Similarly to the case with fixed power allocation, the outage probability $\tilde{\mathrm{P}}_m$ tightly bounds P_m^o. The following lemma provides an expression for the outage probability $\tilde{\mathrm{P}}_m$.

Lemma 3. *When $\rho_I = 0$, the outage probability can be expressed as follows:*

$$\tilde{\mathrm{P}}_m = 1-\Upsilon_1\left(\frac{2\epsilon_{m'}}{\rho}\right)\Upsilon_2\left(\frac{2\epsilon_m(1+\epsilon_{m'})}{\rho}\right), \tag{43}$$

where $\Upsilon_1(y) = \frac{1}{r^2-r_1^2}\left(e^{-yr^\alpha}r^2 - e^{-yr_1^\alpha}r_1^2\right) + \frac{y^{-\frac{2}{\alpha}}}{r^2-r_1^2}\left(\gamma\left(\frac{2}{\alpha}+1, yr^\alpha\right)-\gamma\left(\frac{2}{\alpha}+1, yr_1^\alpha\right)\right)$, *and* $\Upsilon_2(z) = \frac{r_0^2e^{-zr_0^\alpha}}{r_1^2} + \frac{1}{r_1^2}\left(e^{-zr_1^\alpha}r_1^2 - e^{-zr_0^\alpha}r_0^2\right) + \frac{z^{-\frac{2}{\alpha}}}{r_1^2}\left(\gamma\left(\frac{2}{\alpha}+1, zr_1^\alpha\right)-\gamma\left(\frac{2}{\alpha}+1, zr_0^\alpha\right)\right)$. *At high SNR, the outage probability can be approximated as follows:*

$$\begin{aligned}\tilde{\mathrm{P}}_m \approx& \frac{4\epsilon_{m'}}{\rho(2+\alpha)(r^2-r_1^2)}\left(r^{\alpha+2}-r_1^{\alpha+2}\right) \\ &+\frac{2r_0^{2+\alpha}\epsilon_m(1+\epsilon_{m'})}{\rho r_1^2}+\frac{4\epsilon_m(1+\epsilon_{m'})}{\rho(2+\alpha)r_1^2}\left(r_1^{\alpha+2}-r_0^{\alpha+2}\right).\end{aligned} \tag{44}$$

Proof. Please refer to Appendix C. □

By using the above lemma, it is straightforward to show that a diversity gain of one is still achievable at user m (i.e., there is no error floor), and it is important to point out that this is achieved when user m' experiences the same outage performance as if it solely uses the channel. Therefore, by using the proposed cognitive radio NOMA, one additional user, user m, is introduced into the system to share the spectrum with the primary user, user m', without causing any performance degradation at user m'.

Algorithm 1 The selection of the detection vectors $\mathbf{v}_m$ and $\mathbf{v}_{m'}$

1: **for** $i = 1$ to $(2N-M)$ **do**
2: Set $\mathbf{x}_{m,i} = \begin{bmatrix}\mathbf{0}_{1\times(i-1)} & 1 & \mathbf{0}_{1\times(M-i)}\end{bmatrix}^H$, $\forall m \in \{1,\cdots,M\}$.
3: Choose the detection vector as $\begin{bmatrix}\mathbf{v}_{m,i}^H & \mathbf{v}_{m',i}^H\end{bmatrix}^H = \mathbf{U}_m\mathbf{x}_{m,i}$ and determine vector $\mathbf{g}_{m,i} = \mathbf{G}_m^H\mathbf{v}_{m,i}$.
4: Construct the effective small scale fading matrix, denoted by $\bar{\mathbf{G}}_i$, by using $\mathbf{g}_{m,i}$, i.e., $\bar{\mathbf{G}}_i = \begin{bmatrix}\mathbf{g}_{1,i} & \cdots & \mathbf{g}_{M,i}\end{bmatrix}^H$
5: Find the effective small scale fading gain for each user pair, $\gamma_{m,i} = \frac{1}{(\bar{\mathbf{G}}_i^{-1}\bar{\mathbf{G}}_i^{-H})_{m,m}}$.
6: Find the smallest fading gain, $\gamma_{\min,i} = \min\{\gamma_{1,i},\cdots,\gamma_{M,i}\}$.
7: **end for**
8: Find the index i which maximizes the smallest fading gain, $i^* = \underset{i\in\{1,\cdots,2N-M\}}{\arg\max}\ \gamma_{\min,i}$.

C. Selection of the User Detection Vectors

Previously, a random choice of $\mathbf{v}_m$ and $\mathbf{v}_{m'}$ has been used and analyzed. In the case of $2N-M>1$, there is more than one possible choice of the defined vectors in the null space, $\mathbf{U}_m$, defined in (8). In this section, we study how to utilize these additional degrees of freedom and analyze their impact on the outage probability.

Finding the optimal choice for $\mathbf{v}_m$ and $\mathbf{v}_{m'}$ is challenging, since the choice of the detection vectors for one user pair has an impact on those of the other user pairs. For example, the choice of $\mathbf{v}_m$ and $\mathbf{v}_{m'}$ will affect the m-th column of the effective fading matrix $\mathbf{G}$. Recall that the data rates of the users from the i-th pair are a function of $\frac{1}{(\mathbf{G}^{-1}\mathbf{G}^{-H})_{i,i}}$. Therefore, the detection vector chosen by the m-th user pair will also affect the data rates of the users in the i-th pair, $m \neq i$.

In order to avoid this tangled effect, a simple algorithm for detection vector selection is proposed in Table 1. The following lemma shows the diversity gain achieved by the proposed selection algorithm.

Lemma 4. *Consider the use of a fixed set of power allocation coefficients. If $\alpha_{m'}^2 \leq \alpha_m^2\epsilon_{m'}$, the probability $\tilde{\mathrm{P}}_{m'} = 1$, otherwise the use of the algorithm proposed in Table 1 ensures that a diversity gain of $(2N-M)$ is achieved.*

Proof. Please refer to Appendix D. □

As can be seen from Lemma 4, the use of the proposed selection algorithm can increase the diversity gain from 1 to $(2N-M)$, which is a significant improvment compared to the scheme in [15]. Consider a scenario with $N=M$ as an example. The proposed scheme can achieve a diversity gain of M, whereas the one in [15] can only achieve a diversity gain of 1, for an unordered user. Note, however, that the scheme in [15] does not require CSI at the transmitter.

IV. Performance Analysis of MIMO-NOMA Uplink Transmission

Because of the symmetry between the uplink and downlink system models in Section II, in this section, we only focus on the difference between the two cases. One important

observation for uplink NOMA is that the sum rate is always the same, no matter which decoding order is used. Therefore, in this section, we first analyze the outage probability with respect to the sum rate for a fixed power allocation. The use of a randomly selected $\mathbf{x}_m$ is considered in order to obtain tractable analytical results.

A. Fixed Power Allocation

Recall that, if the message from user m is decoded first, the base station can correctly decode the message with rate

$$R_{m,BS,I} = \log\left(1 + \frac{\rho|h_m|^2\alpha_m^2}{\rho|h_{m'}|^2\alpha_{m'}^2 + I_{BS,m} + 1}\right). \quad (45)$$

After subtracting the message from user m, the base station can decode the message from user m' correctly with the following rate $R_{m',BS,I} = \log\left(1 + \frac{\rho|h_{m'}|^2\alpha_{m'}^2}{I_{BS,m}+1}\right)$. Therefore, the sum rate achieved by NOMA in the m-th sub-channel is given by

$$R_s = \log\left(1 + \frac{\rho|h|_m^2\alpha_m^2 + \rho|h|_{m'}^2\alpha_{m'}^2}{I_{BS,m} + 1}\right). \quad (46)$$

It is straightforward to verify that the exactly same sum rate is achieved if the message from user m' is decoded first. Therefore, the outage probability for the sum rate can be expressed as follows:

$$\mathrm{P}_s = \mathrm{P}\left(R_s < R_m + R_{m'}\right). \quad (47)$$

Note that the term for the interference power contains $|\mathbf{p}_m^H\mathbf{1}_M|^2$ which makes the calculation very difficult. Since $|\mathbf{p}_m^H\mathbf{1}_M|^2 \leq M|\mathbf{p}_m^H|^2 = M$, we focus on the following modified expression of the outage probability

$$\tilde{\mathrm{P}}_s = \mathrm{P}\left(\log\left(1 + \frac{\rho|h|_m^2\alpha_m^2 + \rho|h|_{m'}^2\alpha_{m'}^2}{\delta I_m + 1}\right) < R_m + R_{m'}\right), \quad (48)$$

where $I_m = \sum_{j\in\Psi_I} \frac{\rho_I}{L\left(d_{I_j,BS}^{\alpha}\right)}$. Similarly to the downlink case, $\tilde{\mathrm{P}}_s$ provides an upper bound on P_s for $\delta \geq M$. In the simulation section, we will demonstrate that $\tilde{\mathrm{P}}_s$ with a choice of $\delta = 1$ provides a tight approximation to P_s.

Define the small scale fading gain as $x \triangleq \frac{1}{(\mathbf{G}^{-1}\mathbf{G}^{-H})_{m,m}}$. The sum rate outage probability can be expressed as follows

$$\begin{aligned}\tilde{\mathrm{P}}_s &= \mathrm{P}\left(\frac{\rho\frac{x}{L(d_m)}\alpha_m^2 + \rho\frac{x}{L(d_{m'})}\alpha_{m'}^2}{\delta I_m + 1} < \epsilon\right) \quad (49)\\ &= \mathrm{P}\left(x < \frac{\epsilon(\delta I_m + 1)}{\frac{\rho\alpha_m^2}{L(d_m)} + \frac{\rho\alpha_{m'}^2}{L(d_{m'})}}\right),\end{aligned}$$

where $\epsilon = 2^{R_m+R_{m'}} - 1$. Following the same steps as in the proof of Lemma 1, the above probability can be expressed as follows:

$$\begin{aligned}\tilde{\mathrm{P}}_s =& 1 - \frac{4}{r_1^2(r^2 - r_1^2)}\int_{r_1}^{r}\int_0^{r_1} e^{-\zeta(x,y)} \quad (50)\\ &\times e^{-\pi\lambda_I(\rho_I\delta\zeta(x,y))^{\frac{2}{\alpha}}\gamma\left(\frac{1}{\alpha},\frac{\rho_I\delta\zeta(x,y)}{r_0^{\alpha}}\right)} x dx dy dy,\end{aligned}$$

where $\zeta(d_m, d_{m'}) = \frac{\epsilon}{\frac{\rho\alpha_m^2}{L(d_m)} + \frac{\rho\alpha_{m'}^2}{L(d_{m'})}}$.

In order to obtain some insights regarding the above probability, we again consider the case that ρ tends to infinity and ρ_I is fixed. Since both d_m and $d_{m'}$ are bounded, $\zeta(d_m, d_{m'})$ approaches zero at high SNR. Therefore the above probability can be approximated as follows:

$$\begin{aligned}\tilde{\mathrm{P}}_s \approx& 1 - \frac{4}{r_1^2(r^2 - r_1^2)}\int_{r_1}^{r}\int_0^{r_1} e^{-\zeta(x,y)} \quad (51)\\ &\times e^{-\pi\lambda_I(\rho_I\delta\zeta(x,y))^{\frac{2}{\alpha}}\alpha\left(\frac{\rho_I\delta\zeta(x,y)}{r_0^{\alpha}}\right)^{\frac{1}{\alpha}}} x dx dy dy.\end{aligned}$$

With some algebraic manipulations, the above probability can be simplified as follows:

$$\tilde{\mathrm{P}}_s \approx \frac{4\left(\frac{\pi\lambda_I\delta\alpha\rho_I}{r_0} + 1\right)}{r_1^2(r^2 - r_1^2)}\int_{r_1}^{r}\int_0^{r_1}\zeta(x,y)x dx dy dy. \quad (52)$$

Therefore, the outage probability can be approximated as follows:

$$\tilde{\mathrm{P}}_s \approx \frac{4\xi\epsilon\left(\frac{\pi\lambda_I\delta\alpha\rho_I}{r_0} + 1\right)}{\rho r_1^2(r^2 - r_1^2)} \sim \frac{1}{\rho}, \quad (53)$$

where $\xi = \int_{r_1}^{r}\int_0^{r_1}\frac{xy}{\frac{\alpha_m^2}{L(x)} + \frac{\alpha_{m'}^2}{L(y)}} dx dy$ is a constant and not related to the SNR. Hence, a diversity gain of 1 is achievable for the sum rate.

B. Cognitive Radio Power Allocation

The design of cognitive radio NOMA for uplink transmission is more complicated, as explained in the following. To simplify the illustration, we omit the interference term in this section, i.e., $\rho_I = 0$. For downlink transmission, $\alpha_m^2 < \frac{1}{2}$ was sufficient to decide the SIC decoding order. However, there are more uncertainties in the uplink case, since $\alpha_{m'}^2|h_{m'}|^2$ is not necessarily larger than $\alpha_m^2|h_m|^2$ even if $\alpha_{m'}^2 > \frac{1}{2}$. Therefore, the base station can apply two types of decoding strategies, i.e., it may decode the message from user m' first, or that of user m first. These strategies will yield different tradeoffs between the outage performance of the two users, as explained in the following subsections, respectively.

1) Case I: When the message from user m' is decoded first, in order to guarantee the QoS at user m', we impose the following power constraint for the power allocation coefficients

$$\log\left(1 + \frac{\rho|h_{m'}|^2\alpha_{m'}^2}{\rho|h_m|^2\alpha_m^2 + 1}\right) > R_{m'}, \quad (54)$$

which leads to the following choice for $\alpha_{m'}$

$$\alpha_{m'}^2 = \min\left\{1, \frac{\epsilon_{m'} + \rho\epsilon_{m'}|h_m|^2}{\rho|h_{m'}|^2 + \epsilon_{m'}\rho|h_m|^2}\right\}. \quad (55)$$

Following the same steps as in the proof of Lemma 2, the outage probability $\mathrm{P}_{m',BS}^I$ can be evaluated as follows:

$$\begin{aligned}\mathrm{P}_{m',BS}^I &= \mathrm{P}\left(\frac{\epsilon_{m'} + \rho\epsilon_{m'}|h_m|^2}{\rho|h_{m'}|^2 + \epsilon_{m'}\rho|h_m|^2} > 1\right) \quad (56)\\ &= \mathrm{P}\left(|h_{m'}|^2 < \frac{\epsilon_{m'}}{\rho}\right) = 1 - \Upsilon_1\left(\frac{\epsilon_{m'}}{\rho}\right),\end{aligned}$$

and following the same steps as in the proof of Lemma 3, the outage probability $\mathrm{P}^{I}_{m,BS}$ can be evaluated as follows:

$$\begin{aligned}\mathrm{P}^{I}_{m,BS} =&\mathrm{P}\left(|h_{m'}|^2 < \frac{\epsilon_{m'}}{\rho}\right) + \mathrm{P}\left(x > \frac{\epsilon_{m'} L(d_{m'})}{\rho}, \right. \\ & \left. x < L(d_m)\frac{\epsilon_m}{\rho} + L(d_{m'})\frac{\epsilon_{m'}}{\rho}\left(1+\epsilon_m\right)\right) \\ =&1 - \Upsilon_1\left(\frac{\epsilon_{m'}\left(1+\epsilon_m\right)}{\rho}\right)\Upsilon_2\left(\frac{\epsilon_m}{\rho}\right).\end{aligned} \tag{57}$$

2) Case II: When the message from user m is decoded first, in order to guarantee the QoS at user m', we impose the following power constraint for the power allocation coefficients

$$\log\left(1+\rho|h_{m'}|^2\alpha^2_{m'}\right) > R_{m'}, \tag{58}$$

which leads to the following choice for $\alpha_{m'}$

$$\alpha^2_{m'} = \min\left\{1, \frac{\epsilon_{m'}}{\rho|h_{m'}|^2}\right\}. \tag{59}$$

With this choice, we can ensure that the outage probabilities of both users are identical, i.e., $\mathrm{P}^{II}_{m,BS} = \mathrm{P}^{II}_{m',BS}$, as explained in the following. The outage events that occur at user m' can be divided into the following three events

- $\tilde{E}_1$: All the power is allocated to user m', i.e., $\alpha_{m'} = 1$, but the user is still in outage. The NOMA system is degraded to a scenario in which only user m' is served.
- $\tilde{E}_2$: When $\alpha^2_{m'} < 1$, outage occurs at user m, and SIC is stopped.
- $\tilde{E}_3$: When $\alpha^2_{m'} < 1$, no outage occurs at user m, but outage occurs at user m'.

It is straightforward to show that $\tilde{E}_3$ will not happen, i.e., $\mathrm{P}(\tilde{E}_3) = 0$. Therefore $\mathrm{P}_{m',BS} = \mathrm{P}(\tilde{E}_1) + \mathrm{P}(\tilde{E}_2)$. On the other hand, there are only two outage events for decoding the message from user m, which are $\tilde{E}_1$ and $\tilde{E}_2$, respectively. Therefore, the outage probabilities of the two users are the same, $\mathrm{P}^{II}_{m,BS} = \mathrm{P}^{II}_{m',BS}$.

Therefore, we only need to study the outage probability for the message from user m. With the choice shown in (59), the outage probability can be rewritten as follows:

$$\begin{aligned}\mathrm{P}^{II}_{m,BS} =& \mathrm{P}\left(\frac{\epsilon_{m'}}{\rho|h_{m'}|^2} > 1\right) \\ &+ \mathrm{P}\left(\frac{\epsilon_{m'}}{\rho|h_{m'}|^2} < 1, \frac{\rho|h_m|^2\left(1-\frac{\epsilon_{m'}}{\rho|h_{m'}|^2}\right)}{\rho|h_{m'}|^2\frac{\epsilon_{m'}}{\rho|h_{m'}|^2}+1} < \epsilon_m\right).\end{aligned} \tag{60}$$

Therefore, the outage probability can be expressed as follows:

$$\begin{aligned}\mathrm{P}^{II}_{m,BS} =& \mathrm{P}\left(x < \frac{L(d_{m'})\epsilon_{m'}}{\rho}\right) + \mathrm{P}\left(\frac{L(d_{m'})\epsilon_{m'}}{\rho} < \right. \\ & \left. x < \frac{L(d_{m'})\epsilon_{m'}}{\rho} + \frac{\epsilon_m\left(\epsilon_{m'}+1\right)L(d_m)}{\rho}\right).\end{aligned} \tag{61}$$

By applying the same steps as in the proof of Lemma 3 for finding $\mathrm{P}(E_1)$ and $\mathrm{P}(E_3)$, the outage probability can be obtained as follows:

$$\mathrm{P}^{II}_{m',BS} = \mathrm{P}^{II}_{m,BS} = 1 - \Upsilon_1\left(\frac{\epsilon_{m'}}{\rho}\right)\Upsilon_2\left(\frac{\epsilon_m(1+\epsilon_{m'})}{\rho}\right). \tag{62}$$

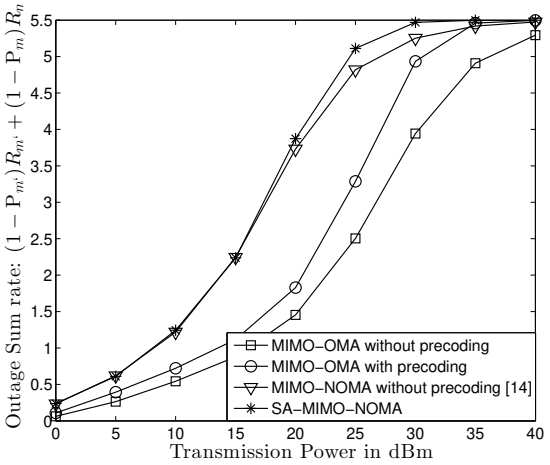
Remark 3: The two considered cases strike different trade-offs between the outage performance of the two users. Case I can ensure that the QoS at user m' is strictly met, and therefore user m' will experience a lower outage probability in Case I, which can be confirmed by the fact that $\mathrm{P}^{I}_{m',BS} < \mathrm{P}^{II}_{m',BS}$, due to $\Upsilon_2\left(\frac{\epsilon_m(1+\epsilon_{m'})}{\rho}\right) \leq 1$. On the other hand, Case II does not require that the message of user m' arrives at the base station with a stronger signal strength since the base station will decode the message from user m first. This is important to avoid the problem of using too much power for compensating the huge path loss of the channel of user m'. As a result, more power is allcoated to user m compared to Case I, and hence, user m experiences better outage performance in Case II, i.e., $\mathrm{P}^{I}_{m,BS} > \mathrm{P}^{II}_{m,BS}$. This can be shown by comparing (57) with (61) and by considering

$$\begin{aligned}&L(d_m)\epsilon_m + \epsilon_{m'}\left(\epsilon_m+1\right)L(d_{m'}) < L(d_{m'})\epsilon_{m'} \\ &+ \epsilon_m\left(\epsilon_{m'}+1\right)L(d_m).\end{aligned} \tag{63}$$

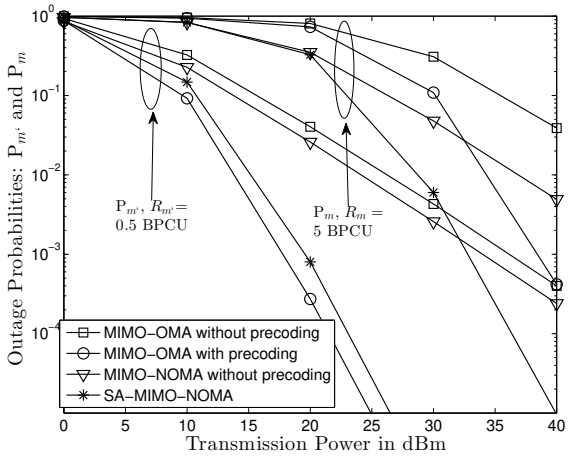
V. Numerical Studies

In this section, the performance of the proposed NOMA framework is investigated by using computer simulations. The performance of three benchmark schemes, termed *MIMO-OMA without precoding*, *MIMO-OMA with precoding*, and *MIMO-NOMA without precoding*, is shown in Fig. 3, in order to better illustrate the performance gain of the proposed framework. The design for the two schemes without precoding can be found in [15]. The MIMO-OMA scheme with precoding serves M users during each orthogonal channel use, e.g., one time slot, whereas $2M$ users are served simultaneously by the proposed scheme. For MIMO-OMA with precoding, the design of the detection vectors was obtained by following the algorithm proposed in Table 1, where the users will carry out antenna selection in each iteration. The framework proposed in this paper is termed *SA-MIMO-NOMA*. The path loss exponent is set as $\alpha = 3$. The size of $\mathcal{D}_1$ and $\mathcal{D}_2$ is determined by $r = 20$m, and $r_1 = 10$m. The parameter for the bounded path loss model is set as $r_0 = 1$.

Since the benchmark schemes were proposed for the interference-free scenario, Fig. 3 shows the performance comparison of the four schemes for $\rho_I = 0$. In Fig. 3(a), the downlink outage sum rate, defined as $R_{m'}(1-\mathrm{P}_{m'}) + R_m(1-\mathrm{P}_m)$, is shown as a function of transmission power, and the corresponding outage probabilities are studied in Fig. 3(b). As can be seen from the figures, the two NOMA schemes can achieve larger outage sum rates compared to the two OMA schemes, which demonstrates the superior spectral efficiency of NOMA. In Fig. 3(b), the two schemes with precoding can achieve better outage performance than the two schemes without precoding, due to the efficient use of the degrees of freedom at the base station. Comparing SA-MIMO-NOMA with the MIMO-NOMA scheme proposed in [15], one can observe that their outage sum rate performances are similar, but SA-MIMO-NOMA can offer much better reception reliability, particularly with high transmission power. In terms of individual outage probability, SA-MIMO-NOMA can ensure a lower outage probability at user m, i.e., a smaller P_m,



(a) Outage Sum rate



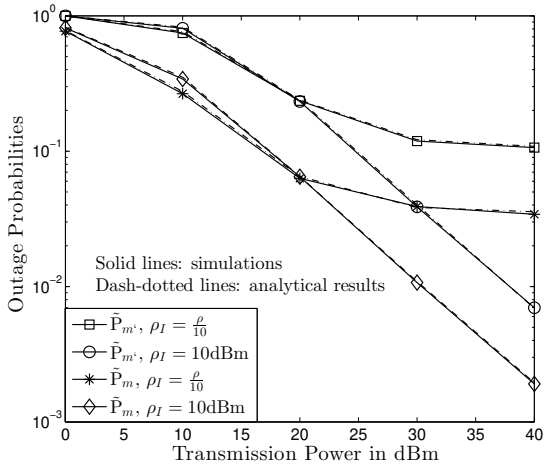
(b) Outage Probabilities

Fig. 3. Performance comparison with the three benchmark schemes for downlink transmission. $R_m = 5$ BPCU and $R_{m'} = 0.5$ BPCU. $r = 20$m and $r_1 = 10$m. $M = N = 3$. $r_0 = 1$m. $a_{m'} = \frac{3}{4}$. The path loss exponent is $\alpha = 3$. The noise power is -30dBm and the interference power is $\rho_I = 0$.

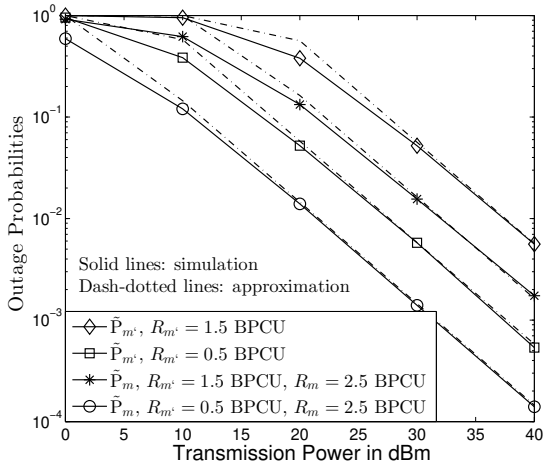
compared to the MIMO-OMA scheme with precoding, but results in performance degradation for the outage probability at user m', i.e., an increase of $\mathrm{P}_{m'}$. This is consistent with the finding in [23] which shows that the NOMA user with poorer channel conditions will suffer some performance loss due to the co-channel interference from its partner.

In Fig. 4, the accuracy of the analytical results developed in Lemmas 1 and 2 for downlink transmission is verified. As can be seen from Fig. 4(a), the exact expression developed in Lemma 1 perfectly matches the computer simulations, and the asymptotic results developed in Lemma 1 are also accurate at high SNR, as shown in Fig. 4(b). The accuracy of Lemma 2 can be confirmed similarly. Note that error floors appear when increasing ρ_I in Fig. 4(a), which is expected due to the strong co-channel interference caused by the randomly deployed interferers.

In Fig. 5, the performance of the cognitive radio power allocation scheme proposed in Section III-B is studied. In particular, given the target data rate at user m', the power allocation coefficients can be calculated opportunistically according to (40). As can be seen from the figure, the probability for this NOMA system to support the secondary user, i.e.,



(a) Exact expressions ($R_{m'} = R_m = 1$ BPCU)



(b) Approximation ($\rho_I = -10$dBm)

Fig. 4. Outage probabilities $\tilde{\mathcal{P}}_{m'}$ and $\tilde{\mathcal{P}}_m$ for downlink transmission. $\lambda_I = 10^{-4}$, $\delta = 1$, $r = 20$m, $r_1 = 10$m, $M = N = 2$, $r_0 = 1$m, and $a_{m'} = \frac{3}{4}$. The path loss exponent is $\alpha = 3$ and the noise power is -30dBm. The analytical results are based on Lemmas 1 and 2.

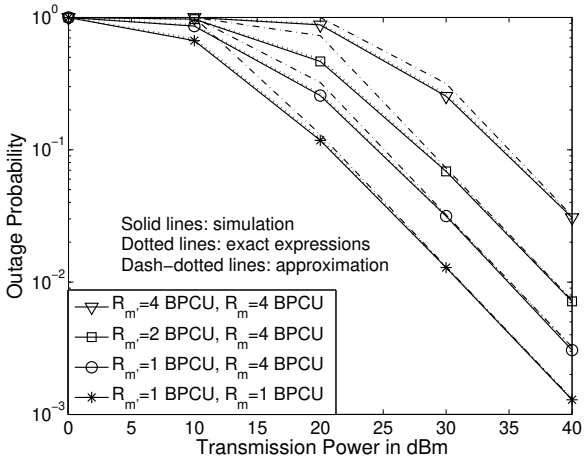


Fig. 5. Outage probability $\tilde{\mathrm{P}}_m$ for cognitive radio downlink transmission. $r = 20$m, $r_1 = 10$m, $r_0 = 1$m, $\delta = 1$, $\rho_I = 0$, and $M = N = 2$. The noise power is -30dBm. The analytical results and the approximations are based on Lemma 3.

user m, with a target data rate of R_m approaches one at high SNR. Note that with OMA, user m cannot be admitted into the channel occupied by user m', and with cognitive radio NOMA, one additional user, user m, can be served without degrading the outage performance of the primary user, i.e., user m'.

In Fig. 6, the impact of the number of user antennas on the outage probability is studied. As can be seen from the figure, by increasing the number of the user antennas, the outage probability is decreased, since the dimension of the null space, $\mathbf{U}_m$, defined in (8), is increased and there are more possible choices for the detection vectors. Furthermore, the slope of the outage curves is also increased, which indicates an increase of the achieved diversity order and hence confirms the findings of Lemma 4.

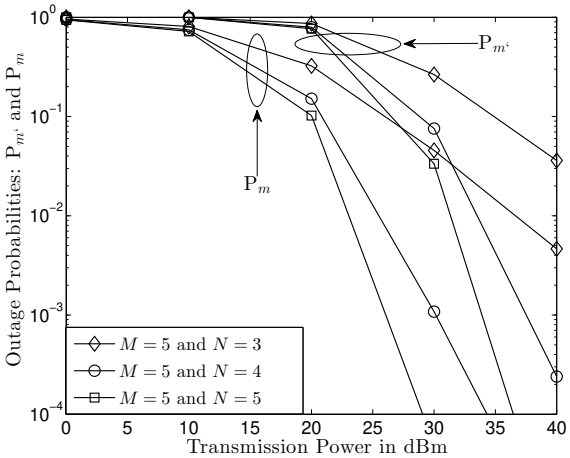


Fig. 6. Impact of the number of the user antennas on the downlink outage probabilities P_m and $\mathrm{P}_{m'}$. $R_m = 4$ BPCU, $R_{m'} = 1.9$ BPCU, $a_{m'} = \frac{3}{4}$, $\rho_I = 0$, $r = 20$m, $r_1 = 10$m, and $r_0 = 1$m. The noise power is -30dBm.

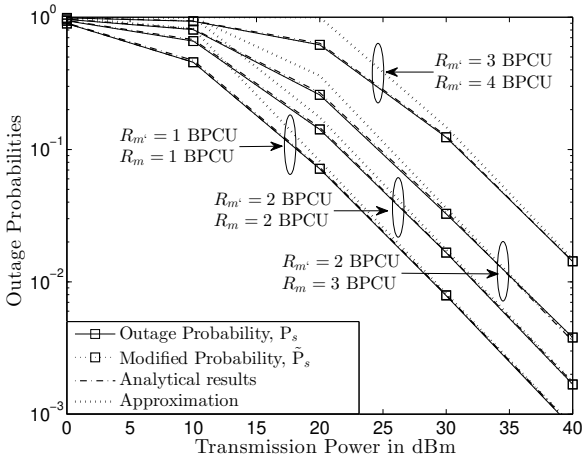


Fig. 7. Outage probabilities for the uplink sum rate P_s and $\tilde{\mathrm{P}}_s$. $r = 20$m, $r_1 = 10$m, $r_0 = 1$m, $\delta = 1$, $\lambda_I = 10^{-4}$, $\rho_I = 10$dB, and $M = N = 2$. The noise power is -30dBm, and the interference power is $\rho_I = -10$dBm. The analytical results and the approximations are based on (50) and (53), respectively.

The performance of the proposed NOMA framework for uplink transmission is demonstrated in Figs. 7 and 8. In particular, in Fig. 7, the outage probability for the sum rate is investigated, and in Fig. 8 the performance of the proposed cognitive radio uplink schemes is studied. As can be observed

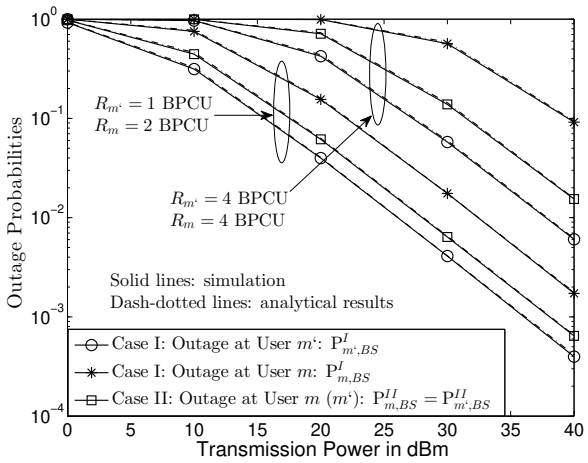


Fig. 8. Uplink outage probability for user m with the cognitive radio constraint. $r = 4$m, $r_1 = 2$m, $r_0 = 1$m, $\rho_I = 0$, and the noise power is -30dBm. The analytical results and the approximations are based on (57) and (62), respectively.

from both figures, the developed analytical results perfectly match the computer simulation results, which demonstrates the accuracy of the developed analytical framework. It is worth pointing out that the modified probability $\tilde{\mathrm{P}}_s$ with $\delta = 1$ provides an accurate approximation for P_s. An interesting observation from Fig. 8 is that Cases I and II offer different performance advantages. In terms of $\mathrm{P}_{m'}$, Case I can offer a lower outage probability compared to Case II, however it results in a loss in outage performance for user m. In practice, if the QoS requirement at user m' is strict, Case I should be used, since the outage probability realized by Case I is exactly the same as when the entire bandwidth is solely occupied by user m'. Otherwise, the use of Case II is more preferable since the outage performance for user m can be improved and the system will not spend exceedingly high powers to compensate the user with poorer channel conditions. One can also observe that, for Case I with $R_{m'} = R_m$, the outage performance for user m is worse than that of user m', although user m is closer to the base station. The reason for this is because in Case I, the power is allocated to user m' first, and user m is served only if there is any power left. Therefore, the outage probability of user m will be at least the same as that of user m', as discussed in Section IV.

VI. Conclusions

In this paper, we have proposed a signal alignment based framework which is applicable to both MIMO-NOMA downlink and uplink transmission. By applying tools from stochastic geometry, the impact of the random locations of the users and interferers has been captured, and closed-form expressions for the outage probability achieved by the proposed framework have been developed to facilitate performance evaluation. In addition to fixed power allocation, a more opportunistic power allocation strategy inspired by cognitive ratio networks has also been investigated. Compared to the existing MIMO-NOMA work, the proposed framework is not only more general, i.e., applicable to both uplink and downlink transmissions, but also offers a significant performance gain in terms of reception reliability.

In this paper, global CSI was assumed to be available for the design of the precoding matrices. It is important to study the design of MIMO-NOMA with imperfect CSI in future work, where one possible solution is to use limited (quantized) CSI feedback in order to reduce the system overhead [30].

APPENDIX A
PROOF FOR LEMMA 1

First, we rewrite the considered probability $\tilde{\mathrm{P}}_{m'}$ as follows:

$$\tilde{\mathrm{P}}_{m'} = \mathrm{P}\left(\frac{\frac{\rho \alpha_{m'}^2}{L(d_{m'})(\mathbf{G}^{-1}\mathbf{G}^{-H})_{m,m}}}{\frac{\rho \alpha_m^2}{L(d_{m'})(\mathbf{G}^{-1}\mathbf{G}^{-H})_{m,m}} + 2 + 2\delta I_{m'}} < \epsilon_{m'} \right). \tag{64}$$

In order to calculate $\tilde{\mathrm{P}}_{m'}$, the density functions for the three parameters, $d_{m'}$, $I_{m'}$, and $\frac{1}{(\mathbf{G}^{-1}\mathbf{G}^{-H})_{m,m}}$ have to be found. Recall from [31] that the factor $\frac{1}{(\mathbf{G}^{-1}\mathbf{G}^{-H})_{m,m}}$ can be written as $\frac{1}{(\mathbf{G}^{-1}\mathbf{G}^{-H})_{m,m}} = \mathbf{g}_m^H (\mathbf{I}_M - \Theta_m)\, \mathbf{g}_m$, where $\Theta_m = \tilde{\mathbf{G}}_m (\tilde{\mathbf{G}}_m^H \tilde{\mathbf{G}}_m)^{-1} \tilde{\mathbf{G}}_m^H$ and $\tilde{\mathbf{G}}_m$ is obtained from $\mathbf{G}$ by removing its m-th row. If $\mathbf{g}_m$ is complex Gaussian distributed, the density function of $\frac{1}{(\mathbf{G}^{-1}\mathbf{G}^{-H})_{m,m}}$ will be exponentially distributed. This can be shown as follows. First, note that the projection matrix $(\mathbf{I}_M - \Theta_m)$ is an idempotent matrix and has eigenvalues which are either zero or one. Second, recall that each row of $\mathbf{G}$ is generated from an $M \times 2N$ complex Gaussian matrix $\begin{bmatrix} \mathbf{G}_m^H & \mathbf{G}_{m'}^H \end{bmatrix}$, i.e.,

$$\mathbf{g}_m = \frac{1}{2} \begin{bmatrix} \mathbf{G}_m^H & \mathbf{G}_{m'}^H \end{bmatrix} \begin{bmatrix} \mathbf{v}_m^H & \mathbf{v}_{m'}^H \end{bmatrix}^H = \frac{1}{2} \begin{bmatrix} \mathbf{G}_m^H & \mathbf{G}_{m'}^H \end{bmatrix} \mathbf{U}_m \mathbf{x}_m. \tag{65}$$

Hence, provided that $\mathbf{x}_m$ is a randomly generated and normalized vector, the application of Proposition 1 in [29] yields the following

$$\mathbf{g}_m \quad \sim \quad \mathrm{CN}(0, \mathbf{I}_M), \tag{66}$$

i.e., $\mathbf{g}_m$ is still an $M \times 1$ complex Gaussian (CN) vector. Therefore, $\frac{1}{(\mathbf{G}^{-1}\mathbf{G}^{-H})_{m,m}}$ is indeed exponentially distributed, and the outage probability can be expressed as follows:

$$\tilde{\mathrm{P}}_{m'} = \mathcal{E}_{I_{m'}, d_{m'}} \left\{ 1 - e^{-2\phi_{m'} L(d_{m'})} \underbrace{e^{-2\delta \phi_{m'} L(d_{m'}) I_{m'}}}_{Q_1} \right\}, \tag{67}$$

which is conditioned on $\alpha_{m'}^2 > \alpha_m^2 \epsilon_{m'}$. Otherwise, $\tilde{\mathrm{P}}_{m'}$ is always one.

Since the homogenous PPP Ψ_I is stationary, the statistics of the interference seen by user m' is the same as that seen by any other receiver, according to Slivnyak's theorem [26]. Therefore, $I_{m'}$ can be equivalently evaluated by focusing on the interference reception seen at a node located at the origin, denoted by $I_0 = \sum_{j \in \Psi_I} \frac{\rho_I}{L\left(d_{I_j}\right)}$, where d_{I_j} denotes the distance between the origin and the j-th interference source. As a result, the expectation of Q_1 with respect to $I_{m'}$ can be expressed as follows: [24], [27]

$$\begin{aligned} \mathcal{E}_{I_{m'}}\{Q_1\} &= \mathcal{E}_{I_{m'}} \left\{ e^{-2\delta \phi_{m'} L\left(d_{m'}\right) \sum\limits_{j \in \Psi_I} \frac{\rho_I}{L\left(d_{I_j}\right)}} \right\} \\ &= \exp\left(-\lambda_I \int_{t \in \mathcal{R}^2} \left(1 - e^{-2\delta \phi_{m'} \rho_I L\left(d_{m'}\right) L(p)} \right) dp \right), \end{aligned} \tag{68}$$

where p denotes the coordinate of the interference source, and d denotes the distance. Note that distance d is determined by the node location p. After changing to polar coordinates, the factor $\mathcal{E}_{I_{m'}}$ can be calculated as follows:

$$\begin{aligned} \mathcal{E}_{I_{m'}}\{Q_1\} &= \exp\left(-\pi \lambda_I r_0^2 \left(1 - e^{-\frac{\beta_{m'}(d_{m'})}{r_0^\alpha}} \right) \right) \\ &\quad \times \exp\left(-2\pi \lambda_I \int_{r_0}^{\infty} \left(1 - e^{-\frac{\beta_{m'}(d_{m'})}{x^\alpha}} \right) x dx \right) \\ &= \exp\left(-\pi \lambda_I \beta_{m'}^{\frac{2}{\alpha}} \gamma\left(\frac{1}{\alpha}, \frac{\beta_{m'}}{r_0^\alpha} \right) \right), \end{aligned} \tag{69}$$

where $\beta_{m'}(d_{m'})$ is denoted by $\beta_{m'}$ for notational simplicity. Therefore, the outage probability can be expressed as follows:

$$\tilde{\mathrm{P}}_{m'} = 1 - \mathcal{E}_{d_{m'}} \left\{ e^{-2\phi_{m'}(d_{m'}^\alpha)} \mathcal{E}_{I_{m'}}\{Q_1\} \right\}. \tag{70}$$

Recall that user m' is uniformly distributed in the ring $\mathcal{D}_2$. Therefore, the above expectation with respect to $d_{m'}$ can be calculated as follows:

$$\tilde{\mathrm{P}}_{m'} = 1 - \int_{p \in \mathcal{D}_2} e^{-2\phi_{m'} L(d_{m'})} \mathcal{E}_{I_{m'}}\{Q_1\} \frac{dp}{\pi r^2 - \pi r_1^2}, \tag{71}$$

where distance $d_{m'}$ is determined by the user location p. Changing again to polar coordinates, this probability can be expressed as follows:

$$\tilde{\mathrm{P}}_{m'} = 1 - \frac{2}{r^2 - r_1^2} \int_{r_1}^{r} e^{-2\phi_{m'} x^\alpha} \mathcal{E}_{I_{m'}}\{Q_1\}\, x dx. \tag{72}$$

Hence, the first part of the lemma is proved.

In the case that ρ approaches infinity and ρ_I is fixed, it is easy to verify that $\phi_{m'}$, as well as $\beta_{m'}$, go to zero. Hence, the incomplete Gamma function in (69) can be approximated as follows:

$$\gamma\left(\frac{1}{\alpha}, \frac{\beta_{m'}}{r_0^\alpha} \right) = \sum_{n=0}^{\infty} \frac{(-1)^n \left(\frac{\beta_{m'}}{r_0^\alpha} \right)^{\frac{1}{\alpha}+n}}{n! \left(\frac{1}{\alpha} + n \right)} \approx \alpha \left(\frac{\beta_{m'}}{r_0^\alpha} \right)^{\frac{1}{\alpha}}. \tag{73}$$

Therefore, the factor $\mathcal{E}_{I_{m'}}$ can be approximated as follows:

$$\mathcal{E}_{I_{m'}}\{Q_1\} \approx \exp\left(-\pi \lambda_I \beta_{m'}^{\frac{2}{\alpha}} \alpha \left(\frac{\beta_{m'}}{r_0^\alpha} \right)^{\frac{1}{\alpha}} \right) \triangleq e^{-d_{m'}^\alpha \theta_{m'}}, \tag{74}$$

where $\theta_{m'} = 2\pi \lambda_I \delta \phi_{m'} \rho_I \frac{\alpha}{r_0}$. Using this approximation the outage probability can be simplified at high SNR as follows:

$$\begin{aligned} \tilde{\mathrm{P}}_{m'} &\approx 1 - \frac{2}{r^2 - r_1^2} \int_{r_1}^{r} e^{-2\phi_{m'} x^\alpha} e^{-x^\alpha \theta_{m'}} x dx \\ &\approx 1 - \frac{2}{r^2 - r_1^2} \int_{r_1}^{r} \left(1 - (2\phi_{m'} + \theta_{m'}) x^\alpha \right) x dx \\ &= \frac{2(2\phi_{m'} + \theta_{m'})}{r^2 - r_1^2} \frac{\left(r^{\alpha+2} - r_1^{\alpha+2} \right)}{\alpha + 2}. \end{aligned} \tag{75}$$

For the special case without co-channel interference, i.e., $\rho_I = 0$, the probability in (72) can be simplified as follows:

$$\begin{aligned} \tilde{\mathrm{P}}_{m'} &= 1 - \frac{2}{r^2 - r_1^2} \int_{r_1}^{r} e^{-2\phi_{m'} x^\alpha} x dx \\ &= 1 - \frac{1}{r^2 - r_1^2} \int_{r_1^\alpha}^{r^\alpha} e^{-2\phi_{m'} y} dy^{\frac{2}{\alpha}}, \end{aligned} \tag{76}$$

and the lemma is proved.

APPENDIX B
PROOF FOR LEMMA 2

When $\alpha_{m'}^2 > \alpha_m^2 \epsilon_{m'}$, the outage probability $\tilde{\mathrm{P}}_m$ can be written as follows:

$$\begin{aligned}\tilde{\mathrm{P}}_m &= \mathrm{P}\left(|h_m|^2 < 2\phi_{m'}(1+\delta I_m)\right) \\ &+ \mathrm{P}\left(|h_m|^2 < 2\phi_m(1+\delta I_m), |h_m|^2 > 2\phi_{m'}(1+\delta I_m)\right) \\ &= \mathrm{P}\left(|h_m|^2 < 2\max\{\phi_m, \phi_{m'}\}(1+\delta I_m)\right),\end{aligned} \tag{77}$$

The reason why $\tilde{\mathrm{P}}_m$ is an upper bound on P_m^o for $\delta \geq N$ can be explained as follows. Recall that the original outage probability P_m^o can be expressed as $\mathrm{P}_m^o = \mathrm{P}\left(|h_m|^2 < \max\{\phi_m, \phi_{m'}\}(|\mathbf{v}_m|^2 + |\mathbf{v}_m^H \mathbf{1}_N|^2 I_m)\right)$. Since $|\mathbf{v}_m^H \mathbf{1}_N|^2 \leq N|\mathbf{v}_m|^2$ and $|\mathbf{v}_m|^2 \leq 2$, we have $\mathrm{P}_m^o \leq \tilde{\mathrm{P}}_m$ if $\delta \geq N$. It is worth pointing out that a choice of $\delta = 1$ is sufficient to yield a tight approximation on P_m^o, as shown in Fig. 2.

Recall that $h_m = \frac{1}{\sqrt{L(d_m)(\mathbf{G}^{-1}\mathbf{G}^{-H})_{m,m}}}$. Comparing h_m to $h_{m'}$, we find that the only difference between the two is the distance d_m which is less than r_1. In addition, the statistics of I_m can be studied by using I_0 as explained in the proof of Lemma 1. Therefore, following steps similar to those in the proof of Lemma 1, the outage probability can be expressed as follows:

$$\tilde{\mathrm{P}}_m = \mathcal{E}_{I_m, d_m}\left\{1 - e^{-2\tilde{\phi}_m L(d_m)} \underbrace{e^{-2\tilde{\phi}_m L(d_m) I_m}}_{Q_2}\right\}. \tag{78}$$

It is straightforward to show that the expectation of Q_2 can be obtained in the same way as that of Q_1, by replacing $\phi_{m'}$ with $\tilde{\phi}_m$. In addition, recall that user m is uniformly distributed in the disc $\mathcal{D}_1$. Therefore, the outage probability can be calculated as follows:

$$\tilde{\mathrm{P}}_m = 1 - \int_{p \in \mathcal{D}_1} e^{-2\tilde{\phi}_m L(d_m)} \varphi_I(L(d_m)) \frac{dp}{\pi r_1^2}, \tag{79}$$

where distance d_m is again determined by the user location p. Resorting to polar coordinates, the outage probability can be expressed as follows:

$$\begin{aligned}\tilde{\mathrm{P}}_m = 1 &- \frac{2}{r_1^2}\int_0^{r_0} e^{-2\tilde{\phi}_m r_0^\alpha} \varphi_I(L(x)) x dx \\ &- \frac{2}{r_1^2}\int_{r_0}^{r_1} e^{-2\tilde{\phi}_m x^\alpha} \varphi_I(L(x)) x dx.\end{aligned} \tag{80}$$

If ρ approaches infinity and ρ_I is fixed, both β_m and $\tilde{\phi}_m$ go to zero. With this approximation, the incomplete Gamma function in (69) can be approximated as $\mathcal{E}_{I_m}\{Q_1\} \approx e^{-d_m^\alpha \theta_m}$, where $\theta_m = 2\pi\lambda_I \tilde{\phi}_m \rho_I \frac{\alpha}{r_0}$. Hence, the outage probability can be simplified at high SNR as follows:

$$\begin{aligned}\tilde{\mathrm{P}}_m &\approx 1 - \frac{2}{r_1^2}\int_0^{r_0} e^{-2\tilde{\phi}_m r_0^\alpha} e^{-r_0^\alpha \theta_m} x dx \\ &- \frac{2}{r_1^2}\int_{r_0}^{r_1} e^{-2\tilde{\phi}_m x^\alpha} e^{-x^\alpha \theta_m} x dx \\ &\approx \frac{(2\tilde{\phi}_m + \theta_m)}{r_1^2(\alpha+2)}\left(\alpha r_0^{\alpha+2} + 2r_1^{\alpha+2}\right),\end{aligned} \tag{81}$$

and the lemma is proved.

APPENDIX C
PROOF FOR LEMMA 3

There are three types of outage events at user m, as illustrated in the following:

- $\bar{\alpha}_m^2 = 0$, i.e., all the power is consumed by user m' and no power is allocated to user m. This event is denoted by E_1.
- When $\bar{\alpha}_m^2 > 0$, user m cannot decode the message to user m'. This event is denoted by E_2.
- When $\bar{\alpha}_m^2 > 0$, user m can decode the message to user m', but fails to decode its own message. This event is denoted by E_3.

The probability of E_1 can be expressed as follows:

$$\mathrm{P}(E_1) = \mathrm{P}\left(\rho|h_{m'}|^2 - 2\epsilon_{m'} < 0\right). \tag{82}$$

This probability can be straightforwardly obtained from the proof of Lemma 1 by replacing $\phi_{m'}$ with $\breve{\phi}_{m'} \triangleq \frac{2\epsilon_{m'}}{\rho}$. Therefore, $\mathrm{P}(E_1)$ can be expressed as follows:

$$\mathrm{P}(E_1) = 1 - \Upsilon_1\left(\frac{2\epsilon_{m'}}{\rho}\right). \tag{83}$$

When $\bar{\alpha}_m > 0$, $\mathrm{P}(E_2) = 0$, since

$$\begin{aligned}\mathrm{P}(E_2) =& \mathrm{P}\left(\frac{\rho|h_m|^2(1-\bar{\alpha}_m^2)}{\rho|h_m|^2\bar{\alpha}_m^2 + 2} < \epsilon_{m'}\right) \\ =& \mathrm{P}\left(\rho|h_m|^2(1-\bar{\alpha}_m^2) < \epsilon_{m'}(\rho|h_m|^2\bar{\alpha}_m^2 + 2)\right) \\ =& \mathrm{P}\left(\rho|h_m|^2 < \bar{\alpha}_m^2\rho|h_m|^2(1+\epsilon_{m'}) + 2\epsilon_{m'}\right) \\ =& \mathrm{P}\left(\rho|h_m|^2|h_{m'}|^2 < \rho|h_{m'}|^2|h_m|^2 - 2|h_m|^2\epsilon_{m'}\right. \\ &\left. + 2|h_{m'}|^2\epsilon_{m'}\right) = \mathrm{P}\left(|h_m|^2 < |h_{m'}|^2\right) = 0.\end{aligned} \tag{84}$$

The probability for event E_3 can be calculated as follows:

$$\begin{aligned}\mathrm{P}(E_3) &= \mathrm{P}\left(\log\left(1 + \frac{\rho}{2}|h_m|^2\alpha_m^2\right) < R_m, \bar{\alpha}_m > 0\right) \\ &= \mathrm{P}\left(|h_m|^2 \frac{\rho|h_{m'}|^2 - 2\epsilon_{m'}}{2(1+\epsilon_{m'})|h_{m'}|^2} < \epsilon_m, |h_{m'}|^2 > \frac{2\epsilon_{m'}}{\rho}\right).\end{aligned} \tag{85}$$

An important observation is that both channel gains h_m and $h_{m'}$ share the same small scale fading. Defining $x = \frac{1}{(\mathbf{G}^{-1}\mathbf{G}^{-H})_{m,m}}$, the outage probability can be expressed as follows:

$$\begin{aligned}\mathrm{P}(E_3) &= \mathrm{P}\left(\frac{x}{L(d_m)} \frac{\rho\frac{x}{L(d_{m'})} - 2\epsilon_{m'}}{(1+\epsilon_{m'})\frac{x}{L(d_{m'})}} < 2\epsilon_m, \frac{x}{L(d_{m'})} > \frac{2\epsilon_{m'}}{\rho}\right) \\ &= \mathrm{P}\left(\frac{2\epsilon_{m'}L(d_{m'})}{\rho} < x < \frac{2\epsilon_{m'}L(d_{m'})}{\rho} + \frac{2\epsilon_m(1+\epsilon_{m'})L(d_m)}{\rho}\right).\end{aligned}$$

The above probability can be calculated as follows

$$\begin{aligned}\mathrm{P}(E_3) =& \int_{p_{m'} \in \mathcal{D}_2} e^{-\frac{2\epsilon_{m'}L(d_{m'})}{\rho}} dp_{m'} \\ &- \int_{p_m \in \mathcal{D}_1, p_{m'} \in \mathcal{D}_2}\int e^{-\frac{2\epsilon_{m'}L(d_{m'})}{\rho} - \frac{2\epsilon_m(1+\epsilon_{m'})L(d_m)}{\rho}} dp_m dp_{m'},\end{aligned} \tag{86}$$

where p_m denotes the location of user m. Since the users are uniformly distributed, the above probability can be expressed as follows:

$$\begin{aligned}\mathrm{P}(E_3) &= \frac{2}{(r^2-r_1^2)}\int_{r_1}^{r} e^{-\frac{2\epsilon_{m'}}{\rho y^{\alpha}}} y dy - \frac{4}{r_1^2(r^2-r_1^2)} \\ &\times \int_0^{r_1} e^{-\frac{2\epsilon_m(1+\epsilon_{m'})}{\rho y^{\alpha}}} y dy \int_{r_1}^{r} e^{-\frac{2\epsilon_{m'}L(x)}{\rho}} x dx \\ &= \Upsilon_1\left(\frac{2\epsilon_{m'}}{\rho}\right) - \Upsilon_1\left(\frac{2\epsilon_{m'}}{\rho}\right)\Upsilon_2\left(\frac{2\epsilon_m(1+\epsilon_{m'})}{\rho}\right).\end{aligned} \tag{87}$$

Combining (83), (84), and (87), the first part of the lemma can be proved. To obtain the high SNR approximation, we have

$$\begin{aligned}\Upsilon_1(y) &\approx 1 + \frac{1}{r^2-r_1^2}\left(y r_1^{\alpha+2} - y r^{\alpha+2}\right) \\ &+ \frac{y^{-\frac{2}{\alpha}}}{(\frac{2}{\alpha}+1)(r^2-r_1^2)}\left(\left(y r^{\alpha}\right)^{\frac{2}{\alpha}+1} - \left(y r_1^{\alpha}\right)^{\frac{2}{\alpha}+1}\right) \\ &= 1 - \frac{2y}{(2+\alpha)(r^2-r_1^2)}\left(r^{\alpha+2} - r_1^{\alpha+2}\right),\end{aligned} \tag{88}$$

when y approaches zero, and

$$\begin{aligned}\Upsilon_2(z) &\approx 1 - \frac{r_0^{2+\alpha}z}{r_1^2} - \frac{1}{r_1^2}\left(z r_1^{\alpha+2} - z r_0^{\alpha+2}\right) \\ &+ \frac{z^{-\frac{2}{\alpha}}}{(\frac{2}{\alpha}+1)r_1^2}\left(\left(z r_1^{\alpha}\right)^{\frac{2}{\alpha}+1} - \left(z r_0^{\alpha}\right)^{\frac{2}{\alpha}+1}\right) \\ &= 1 - \frac{r_0^{2+\alpha}z}{r_1^2} - \frac{2z}{(2+\alpha)r_1^2}\left(r_1^{\alpha+2} - r_0^{\alpha+2}\right),\end{aligned} \tag{89}$$

when z approaches zero. By substituting the approximations into (43), the lemma is proved.

Appendix D
Proof for Lemma 4

We focus on the outage performance of user m' first. Given the detection vector $\mathbf{v}_{m,i^*}$ chosen from Table 1, the outage probability can be upper bounded as follows:

$$\begin{aligned}\mathrm{P}_{m',i^*} &\le \mathrm{P}\left(\frac{\frac{\rho\alpha_{m'}^2}{L(d_{m'})(\bar{\mathbf{G}}_{i^*}^{-1}\bar{\mathbf{G}}_{i^*}^{-H})_{m,m}}}{\frac{\rho\alpha_m^2}{L(d_{m'})(\bar{\mathbf{G}}_{i^*}^{-1}\bar{\mathbf{G}}_{i^*}^{-H})_{m,m}} + 2 + 2\delta I_{m'}} < \epsilon_{m'}\right) \\ &= \mathrm{P}\left(\gamma_{m,i^*} < 2\phi_{m'}L(d_{m'})(1+\delta I_{m'})\right) \le \mathrm{P}\left(\gamma_{\min,i^*}\right. \\ &\left.< 2\phi_{m'}L(d_{m'})(1+\delta I_{m'})\right).\end{aligned} \tag{90}$$

According to the algorithm proposed in Table 1,

$$\gamma_{\min,i^*} = \max\{\gamma_{\min,1}, \cdots, \gamma_{\min,2N-M}\}. \tag{91}$$

Therefore, the outage probability can be bounded as follows:

$$\mathrm{P}_{m',i^*} \le \left(\mathrm{P}\left(\gamma_{\min,i} < 2\phi_{m'}L(d_{m'})(1+\delta I_{m'})\right)\right)^{2N-M},$$

where the inequality follows from the fact that $\gamma_{\min,i}$ and $\gamma_{\min,j}$ are independent, since $\mathbf{g}_{m,i}$ and $\mathbf{g}_{m,j}$ are independent (Proposition 1 in [29]). The above outage probability can be further bounded as follows:

$$\mathrm{P}_{m',i^*} \le \left(M\mathrm{P}\left(\gamma_{m,i} < 2\phi_{m'}L(d_{m'})(1+\delta I_{m'})\right)\right)^{2N-M}. \tag{92}$$

Following the same steps as in the proof of Lemma 1, the upper bound on the outage probability can be calculated as follows:

$$\begin{aligned}\mathrm{P}_{m',i^*} &\le M^{2N-M}\mathcal{E}_{I_{m'},d_{m'}}\left\{\left(1 - e^{-2\phi_{m'}L(d_{m'})}\right.\right. \\ &\left.\left.\times e^{-2\delta\phi_{m'}L(d_{m'})I_{m'}}\right)^{2N-M}\right\} \\ &\le M^{2N-M}\sum_{i=0}^{2N-M}\binom{2N-M}{i}(-1)^i \\ &\times \mathcal{E}_{I_{m'},d_{m'}}\left\{e^{-2i\phi_{m'}L(d_{m'})}e^{-2i\delta\phi_{m'}L(d_{m'})I_{m'}}\right\},\end{aligned} \tag{93}$$

which is conditioned on $\alpha_{m'}^2 > \alpha_m^2\epsilon_{m'}$.

After the expectation with respect to $I_{m'}$, the outage probability can be bounded as follows:

$$\begin{aligned}\mathrm{P}_{m',i^*} &\le M^{2N-M}\sum_{i=0}^{2N-M}\binom{2N-M}{i}(-1)^i \\ &\times \mathcal{E}_{d_{m'}}\left\{e^{-2i\phi_{m'}L(d_{m'})}e^{-\pi\lambda_I(i\beta_{m'})^{\frac{2}{\alpha}}\gamma\left(\frac{1}{\alpha},\frac{i\beta_{m'}}{r_0^{\alpha}}\right)}\right\}.\end{aligned} \tag{94}$$

For the case of ρ approaching infinity and a fixed ρ_I, the upper bound on the outage probability can be approximated as follows:

$$\begin{aligned}\mathrm{P}_{m',i^*} &\le M^{2N-M}\sum_{i=0}^{2N-M}\binom{2N-M}{i}(-1)^i \\ &\times \mathcal{E}_{d_{m'}}\left\{e^{-2i\phi_{m'}L(d_{m'})}e^{-\pi\lambda_I(i\beta_{m'})^{\frac{2}{\alpha}}\alpha\left(\frac{i\beta_{m'}}{r_0^{\alpha}}\right)^{\frac{1}{\alpha}}}\right\} \\ &\le M^{2N-M}\sum_{i=0}^{2N-M}\binom{2N-M}{i}(-1)^i \\ &\times \mathcal{E}_{d_{m'}}\left\{e^{-(i\theta_{m'}+2i\phi_{m'})d_m^{\alpha}}\right\}.\end{aligned} \tag{95}$$

Using polar coordinates, the upper bound can be calculated as follows:

$$\begin{aligned}\mathrm{P}_{m',i^*} &\le M^{2N-M}\sum_{i=0}^{2N-M}\binom{2N-M}{i}(-1)^i\frac{2}{r^2-r_1^2} \\ &\times \sum_{j=0}^{\infty}\int_{r_1}^{r}\frac{(-1)^j(i\theta_{m'}+2i\phi_{m'})^j x^{j\alpha}}{j!}x dx\end{aligned} \tag{96}$$

$$\begin{aligned}&= \frac{2M^{2N-M}}{r^2-r_1^2}\sum_{i=0}^{2N-M}\binom{2N-M}{i}(-1)^i \\ &\times \sum_{j=0}^{\infty}\frac{(-1)^j(i\theta_{m'}+2i\phi_{m'})^j}{j!}\frac{\left(r^{j\alpha+2}-r_1^{j\alpha+2}\right)}{j\alpha+2}.\end{aligned} \tag{97}$$

By exchanging the two sums in the above equation, the

upper bound can be rewritten as follows:

$$\mathrm{P}_{m',i^*} \leq \frac{2M^{2N-M}}{r^2-r_1^2}\sum_{j=0}^{\infty}\frac{(-1)^j(\theta_{m'}+2\phi_{m'})^j}{j!} \tag{98}$$
$$\times \frac{\left(r^{j\alpha+2}-r_1^{j\alpha+2}\right)}{j\alpha+2}\sum_{i=0}^{2N-M}\binom{2N-M}{i}(-1)^i i^j$$
$$= \frac{2M^{2N-M}}{r^2-r_1^2}\sum_{j=2N-M}^{\infty}\frac{(-1)^j(\theta_{m'}+2\phi_{m'})^j}{j!} \tag{99}$$
$$\times \frac{\left(r^{j\alpha+2}-r_1^{j\alpha+2}\right)}{j\alpha+2}\sum_{i=0}^{2N-M}\binom{2N-M}{i}(-1)^i i^j,$$

where the last step follows from the following fact

$$\sum_{i=0}^{2N-M}\binom{2N-M}{i}(-1)^i i^j = 0$$

for $0 \leq j \leq (2N-M-1)$ [32]. Furthermore, note that both $\phi_{m'}$ and $\theta_{m'}$ approach zero for the considered scenario, and $\sum_{i=0}^{2N-M}\binom{2N-M}{i}(-1)^i i^{2N-M} = (-1)^{2N-M}(2N-M)!$. Therefore, the upper bound on the outage probability can be approximated as follows:

$$\mathrm{P}_{m',i^*} \leq \frac{2M^{2N-M}}{r^2-r_1^2}\frac{(-1)^{2N-M}(\theta_{m'}+2\phi_{m'})^{2N-M}}{(2N-M)!}$$
$$\times \frac{\left(r^{(2N-M)\alpha+2}-r_1^{(2N-M)\alpha+2}\right)}{(2N-M)\alpha+2}(-1)^{2N-M}(2N-M)!$$
$$= \frac{2[M(\theta_{m'}+2\phi_{m'})]^{2N-M}\left(r^{(2N-M)\alpha+2}-r_1^{(2N-M)\alpha+2}\right)}{(r^2-r_1^2)((2N-M)\alpha+2)}$$
$$\sim \frac{1}{\rho^{2N-M}}. \tag{100}$$

The result for user m can be proved using steps similar to the ones above.

The result for a random detection vector can be obtained by replacing $(2N-M)$ with 1 in the above expression, and the corresponding upper bound becomes

$$\mathrm{P}_{m',i^*} \leq \frac{2M[\theta_{m'}+2\phi_{m'}]\left(r^{\alpha+2}-r_1^{\alpha+2}\right)}{(r^2-r_1^2)(\alpha+2)}, \tag{101}$$

which is exactly the same result as the one shown in Lemma 1, except for the extra term M which was introduced by upper bounding the outage probability in (92). Hence, the proof is completed.

REFERENCES

[1] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE Vehicular Technology Conference*, Dresden, Germany, Jun. 2013.

[2] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE Annual Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, London, UK, Sept. 2013.

[3] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Letters*, vol. 21, no. 12, pp. 1501–1505, Dec 2014.

[4] 3rd Generation Partnership Project(3GPP), "Study on downlink multiuser superposition transmission for LTE," Mar. 2015.

[5] "5G radio access: requirements, concepts and technologies," NTT DOCOMO, Inc., Tokyo, Japan, 5G Whitepaper, Jul. 2014.

[6] "Proposed solutions for new radio access," Mobile and wireless communications Enablers for the Twenty-twenty Information Society (METIS), Deliverable D.2.4, Feb. 2015.

[7] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Letters*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.

[8] V. Kalokidou, O. Johnson, and R. Piechocki, "A hybrid TIM-NOMA scheme for the SISO broadcast channel," in *Proc. IEEE International Conference on Commun. (ICC)*, Jun. 2015.

[9] ——, "A hybrid TIM-NOMA scheme for the broadcast channel," *EAI Endorsed Transactions on Wireless Spectrum*, vol. 1, no. 3, pp. 1–12, Jul. 2015.

[10] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Uplink non-orthogonal multiple access for 5G wireless networks," in *Proc. 11th International Symposium on Wireless Communications Systems (ISWCS)*, Barcelona, Spain, Aug 2014, pp. 781–785.

[11] H. Marshoud, V. M. Kapinas, G. K. Karagiannidis, and S. Muhaidat, "Non-orthogonal multiple access for visible light communications," *arXiv preprint*, Available on-line at arxiv.org/abs/1504.00934.

[12] J. Choi, "Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems," *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 791–800, Mar. 2015.

[13] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannidis, "A minorization-maximization method for optimizing sum rate in non-orthogonal multiple access systems," *IEEE Trans. Signal Process.*, (to appear in 2015) Available on-line at arxiv.org/abs/1505.05735.

[14] Q. Sun, S. Han, C.-L. I, and Z. Pan, "On the ergodic capacity of MIMO NOMA systems," *IEEE Wireless Commun. Letters*, (to appear in 2015).

[15] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, (to appear in 2015) Available on-line at arXiv:1503.05367.

[16] K. Higuchi and Y. Kishiyama, "Non-orthogonal access with random beamforming and intra-beam SIC for cellular MIMO downlink," in *Proc. IEEE Vehicular Technology Conference*, Las Vegas, NV, US, Sept. 2013, pp. 1–5.

[17] V. R. Cadambe and S. A. Jafar, "Interference alignment and the degrees of freedom for the K user interference channel," *IEEE Trans. Inform. Theory*, vol. 54, pp. 3425–3441, Aug. 2008.

[18] M. Maddah-Ali, A. Motahari, and A. Khandani, "Communication over MIMO X channels: Interference alignment, decomposition, and performance analysis," *IEEE Trans. Inform. Theory*, vol. 54, no. 8, pp. 3457–3470, Aug. 2008.

[19] N. Lee, J.-B. Lim, and J. Chun, "Degrees of the freedom of the MIMO Y channel: Signal space alignment for network coding," *IEEE Trans Inform. Theory*, vol. 56, pp. 3332 – 3343, Jul. 2010.

[20] Z. Ding and H. Poor, "A general framework of precoding design for multiple two-way relaying communications," *IEEE Trans. Signal Process.*, vol. 61, no. 6, pp. 1531–1535, Mar. 2013.

[21] M. Peng, C. Wang, J. Li, H. Xiang, and V. Lau, "Recent advances in underlay heterogeneous networks: Interference control, resource allocation, and self-organization," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 700–729, Second quarter 2015.

[22] "C-RAN: The road towards green RAN," China Mobile Res. Inst., Beijing, China, Oct. 2011, White Paper, ver. 2.5.

[23] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G non-orthogonal multiple access," *IEEE Trans. Vehicular Technology*, (to appear in 2015) Available on-line at arXiv:1412.2799.

[24] M. Haenggi, *Stochastic Geometry for Wireless Networks*. Cambridge University Press, Cambridge, UK, 2012.

[25] J. Venkataraman, M. Haenggi, and O. Collins, "Shot noise models for outage and throughput analyses in wireless ad hoc networks," in *Proc. IEEE Military Communications Conference*, Washington, DC, USA, Oct. 2006.

[26] C.-H. Liu and J. Andrews, "Multicast outage probability and transmission capacity of multihop wireless networks," *IEEE Trans. Inform. Theory*, vol. 57, no. 7, pp. 4344–4358, Jul. 2011.

[27] I. Krikidis, "Simultaneous information and energy transfer in large-scale networks with/without relaying," *IEEE Trans. Commun.*, vol. 62, no. 3, pp. 900–912, Mar. 2014.

[28] N. Lee and J.-B. Lim, "A novel signaling for communication on MIMO Y channel: Signal space alignment for network coding," in *Proc. IEEE International Symposium on Inform. Theory (ISIT-09)*, Jul. 2009.

[29] Z. Ding, T. Wang, M. Peng, and W. Wang, "On the design of network coding for multiple two-way relaying channels," *IEEE Trans. Wireless Commun.*, vol. 10, no. 6, pp. 1820–1832, Jun. 2011.

[30] Z. Ding and H. V. Poor, "Design of massive-MIMO-NOMA with limited feedback," *IEEE Signal Process. Letters*, (submitted) Available on-line at arXiv:1511.05583.
[31] M. Rupp, C. Mecklenbrauker, and G. Gritsch, "High diversity with simple space time block codes and linear receivers," *Proc. IEEE Global Commun. Conf.*, vol. 2, pp. 302–306, San Francisco, USA, Dec. 2003.
[32] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, 6th ed. New York: Academic Press, 2000.

Zhiguo Ding (S'03-M'05) received his B.Eng in Electrical Engineering from the Beijing University of Posts and Telecommunications in 2000, and the Ph.D degree in Electrical Engineering from Imperial College London in 2005. From Jul. 2005 to Aug. 2014, he was working in Queen's University Belfast, Imperial College and Newcastle University. Since Sept. 2014, he has been with Lancaster University as a Chair Professor. From Oct. 2012 to Sept. 2016, he has been also with Princeton University as an Academic Visitor.

Dr Ding' research interests are 5G networks, game theory, cooperative and energy harvesting networks and statistical signal processing. He is serving as an Editor for *IEEE Transactions on Communications*, *IEEE Transactions on Vehicular Networks*, *IEEE Wireless Communication Letters*, *IEEE Communication Letters*, and *Journal of Wireless Communications and Mobile Computing*. He received the best paper award in IET Comm. Conf. on Wireless, Mobile and Computing, 2009, IEEE Communication Letter Exemplary Reviewer 2012, and the EU Marie Curie Fellowship 2012-2014.

Robert Schober (S'98, M'01, SM'08, F'10) was born in Neuendettelsau, Germany, in 1971. He received the Diplom (Univ.) and the Ph.D. degrees in electrical engineering from the University of Erlangen-Nuermberg in 1997 and 2000, respectively. From May 2001 to April 2002 he was a Postdoctoral Fellow at the University of Toronto, Canada, sponsored by the German Academic Exchange Service (DAAD). Since May 2002 he has been with the University of British Columbia (UBC), Vancouver, Canada, where he is now a Full Professor. Since January 2012 he is an Alexander von Humboldt Professor and the Chair for Digital Communication at the Friedrich Alexander University (FAU), Erlangen, Germany. His research interests fall into the broad areas of Communication Theory, Wireless Communications, and Statistical Signal Processing.

Dr. Schober received several awards for his work including the 2002 Heinz MaierLeibnitz Award of the German Science Foundation (DFG), the 2004 Innovations Award of the Vodafone Foundation for Research in Mobile Communications, the 2006 UBC Killam Research Prize, the 2007 Wilhelm Friedrich Bessel Research Award of the Alexander von Humboldt Foundation, the 2008 Charles McDowell Award for Excellence in Research from UBC, a 2011 Alexander von Humboldt Professorship, and a 2012 NSERC E.W.R. Steacie Fellowship. In addition, he received best paper awards from the German Information Technology Society (ITG), the European Association for Signal, Speech and Image Processing (EURASIP), IEEE WCNC 2012, IEEE Globecom 2011, IEEE ICUWB 2006, the International Zurich Seminar on Broadband Communications, and European Wireless 2000. Dr. Schober is a Fellow of the Canadian Academy of Engineering and a Fellow of the Engineering Institute of Canada. From 2012 to 2015, he was the Editor-in-Chief of the *IEEE Transactions on Communications*. He currently serves as a Member-at-Large on the Board of Governors of the IEEE Communication Society and as Chair of the *IEEE Transactions on Molecular, Biological and Multiscale Communications*.

H. Vincent Poor received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 until 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990 he has been on the faculty at Princeton, where he is the Michael Henry Strater University Professor and Dean of the School of Engineering and Applied Science. He has also held visiting appointments at several universities, including most recently at Stanford and Imperial College. His research interests are in the area of wireless networks and related fields. Among his publications in these areas is the recent book Mechanisms and Games for Dynamic Spectrum Allocation (Cambridge University Press, 2014).

Dr. Poor is a Member of the National Academy of Engineering and the National Academy of Sciences, and is a Foreign Member of the Royal Society. He is also a Fellow of the American Academy of Arts and Sciences, the National Academy of Inventors, and other national and international academies. He received the Marconi and Armstrong Awards of the IEEE Communications Society in 2007 and 2009, respectively. Recent recognition of his work includes the 2014 URSI Booker Gold Medal, the 2015 EURASIP Athanasios Papoulis Award, the 2016 John Fritz Medal, and honorary doctorates from Aalborg University, Aalto University, HKUST and the University of Edinburgh.